Mining your Business in Retail with IBM DB2 Intelligent Miner

Presented by DB2 Developer Domain

http://www7b.software.ibm.com/dmdd/

Table of Contents

If you're viewing this document online, you can click any of the topics below to link directly to that section.

Section 1. Introducing this tutorial

Objectives

This tutorial shows you how to apply data mining techniques using IBM DB2 Intelligent Miner to generate automatically product recommendations for customers in a possible e-commerce shop environment. Moreover, we can characterize the customers in terms of the products they purchase and the frequency of their shoppings. This customer profiling will be key to determine targets for an outbound cross-sell campaign. We will be able to ensure that the recommendations will be appropriate to each customer because other customers with a similar behaviour bought what we are now recommending.

To fulfill these objectives, we will show you how to use IBM DB2 Intelligent Miner Modeling to create a mining model (Glossary on page 63) and then the IBM DB2 Intelligent Miner Visualization to evaluate it and to display its results. Furthermore, we will give some examples of how we could integrate all with the existing business applications, as the integration of the data mining results with channels and reporting is key for a successful marketing strategy.

This tutorial takes you through the entire process of extracting product recommendations and customer's profiling step by step, using real data taken from the retail business. By following the examples, you can modify various parameters to observe the resulting outcome. By completing the exercises, you can check what you have learned.

When you have completed this tutorial, you can:

- Understand the concepts of data mining:

1. What is data mining?
2. What can you achieve with data mining?
3. How does the data mining process look like?

- Apply the techniques that you have learned to your own corporate problems and put each of the theoretical steps of the Data Mining process into practice.
- Generate business explanations and deploy them.

NOTE: In the future we may refer to the IBM DB2 Intelligent Miner tools as IM Visualization/IM modeling/ IM Scoring, where IM means Intelligent Miner.

Who should read this tutorial?

This tutorial is addressed to Information Technology (IT) professionals. To complete this tutorial, you should be familiar with databases. You need not be familiar with data mining concepts. If you do not have domain knowledge of the business area, you should work together with people that can provide this knowledge.

Knowing the business area is mandatory. With data mining, you can solve business problems by analyzing the Corporate Data. However, you will not come up with a proper answer or explanation if there is not a concise definition and understanding of the business problem.

Tutorial Scenario

Nowadays internet shops offer product recommendations every time the customer adds a product into the cart. In an early future, there will be little computer devices in each cart of a physical supermarket. Each time a customer puts a new product in the cart, its code will be scanned and a display will show some product recommendations in real time, taking into account the kind of customer that is purchasing and the item he/she has put in the cart.

Right now companies can apply product recommendation to improve their cross-selling campaigns, for example. The cross-selling strategy is based in pushing new products to current customers based on their past purchases. Cross-selling is designed to widen the customer's reliance on the company and decrease the likelihood of the customer switching to a competitor. Indeed, companies are worried in losing their clients as it is very expensive to attract new ones with promotion campaigns, publicity, etc.

This is our scenario:

"A retail company wants to have a better knowledge of their clients and their behaviour so that it can offer a better product recommendation in its next cross selling campaign."

Required Software

To complete this tutorial, the following software must be installed on your computer:

- IBM DB2 UDB Version 8.1 or higher (download *trial version*)
- IBM DB2 Intelligent Miner Modeling Version 8.1 or higher (download *trial version*)
- IBM DB2 Intelligent Miner Visualization Version 8.1 or higher (download *trial version*)
- IBM DB2 Intelligent Miner Scoring Version 8.1 or higher (download *trial version*)

As operating systems, this tutorial supports Windows and Linux.

Duration

To complete this tutorial, you need about 3 to 3.5 hours in average.

However, you might want to read this tutorial more than once when you want to translate the data mining techniques into your own business environment.

Navigating through the tutorial

Navigation links:

- To move forward or backward through the tutorial, click **Next** or **Previous**.
- To continue with the next section, click **Next Section** on the last page of the current section.
- To display the contents of the current section, click **Section menu**.
- To return to the main menu, click **Main menu**.

Cross-reference links:

- To send your feedback, or to address a question to the author about the content of this tutorial, click **Feedback**.
- The links that lead to a Web site open a new browser window.
- The links that lead to different topics within this tutorial, for example, **Troubleshooting**, display the appropriate section of this tutorial.
- To go back to the page where the link started,

  click this button 

  , which will be displayed. You can also click**Back** on the toolbar of your browser. If you click **Previous**, the previous page in the sequence of the page where the link lead to is displayed.

Preparing your computer

Retrieving the required files

The data from the retail company is stored in the file *retail-case* folder.

Save the file IDMEasyMiningProcedures.jar in the *\Program Files\IBM\Sqllib\function* folder (Windows) or the */home/db2inst/sqllib/function* folder (Linux).

Creating and configuring the database and importing the data

Before you can build or call a model, you must create and configure the retail database, and import the data.

Windows:

1. To display a DB2 command window, click **Start** -> **Programs** -> **IBM DB2** -> **Command Line Tools** -> **Command Window**.
2. Go to the directory where you have saved the retail-case folder, for example, :e\retail_tutorial\retail-case\, and type the following commands on the command line:

```
[tutorial path]:\> setup1

[tutorial path]:\> setup2
```

Linux:

1. Start the DB2 opening a console and switching user to a db2 instance. Then execute a **db2start**.
2. To run DB2 Java routines (stored procedures and UDFs), update the DB2 database

manager configuration on the server to include the path where the JDK is installed on the machine. You can do this by entering the the following command:

```
/[tutorial path]> db2 update dbm cfg using
                          JDK_PATH JAVAHOME
```

Where */home/db2inst/jdk13* is the path where the JDK is installed.

3. To run Java stored procedures or user-defined functions, the Linux run-time linker must be able to access certain Java shared libraries, and DB2 must be able to load these libraries and the Java virtual machine. Go to the */usr/lib* directory as *root* and execute:

```
/[tutorial path]> ln -fs JAVAHOME/jre/bin/libjava.so

/[tutorial path]> ln -fs JAVAHOME/jre/bin/classic/libjvm.so

/[tutorial path]> ln -fs JAVAHOME/jre/bin/libhpi.so
```

   Where *JAVAHOME* is the base directory for the JDK.

4. Go to the directory where you have saved the retail-case folder, for example, /retail_tutorial/retail-case/, and type the following commands on the command line:

```
/[tutorial path]> sh setup1.sh
/[tutorial path]> sh setup2.sh
```

The setup1 command creates the database and configures it for modeling and declare the Easy Mining procedures that are used to build and apply the mining models, whereas the setup2 imports the data.

About the authors: M. Eugenia Garcia Bordes

M. Eugenia Garcia Bordes is a Computer Science student at the Facultat de Informatica de Barcelona (*FIB*), which belongs to the Universitat Politecnica de Catalunya (*UPC*) in Spain. Despite her technical background, she has attended several courses about management and economics while her studies.

On the other hand, she joined the *Unitech* International exchange programme in 2002-2003. During an intensive, full year abroad, thanks to this programme Eugenia complemented her engineering studies with international academic and industry exposure, joint courses and insights in management theory and practise. Within this framework, after studying one semester at Technische Universiteit Delft (*TUDelft*) in the Netherlands, she spent half a year in the *IBM* Research and Development Laboratory in Boeblingen (Germany), where she developed this tutorial.

About the authors: Toni Bollinger

Toni Bollinger studied Computer Science in Bonn, Germany, and Paris, France. He joined IBM in 1987 where he worked in several AI research projects. Since 1994 he works on data mining in the development of the Intelligent Miner software products as well as in practical data mining projects. His main focus in his current activities is to make data mining easier and accessible to a wider range of users.

You can contact him at IBM under *Toni.Bollinger@de.ibm.com*

Trademarks

The following terms are trademarks of the IBM Corporation in the United States, other countries, or both:

- DB2
- DB2 Universal Database
- Websphere
- IBM
- Intelligent Miner
- Redbooks

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

Java and all Java-based trademarks are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.

Other company, product, and service names may be trademarks or service marks of others.

Section 2. Introducing data mining

What is data mining?

"Data Mining is the process of extracting valid, useful, unknown and comprehensible information from data and using it to make business decisions."

The key issue in this definition is that the information that Data Mining tools are able to generate is *useful* and *unknown*.

Reference:

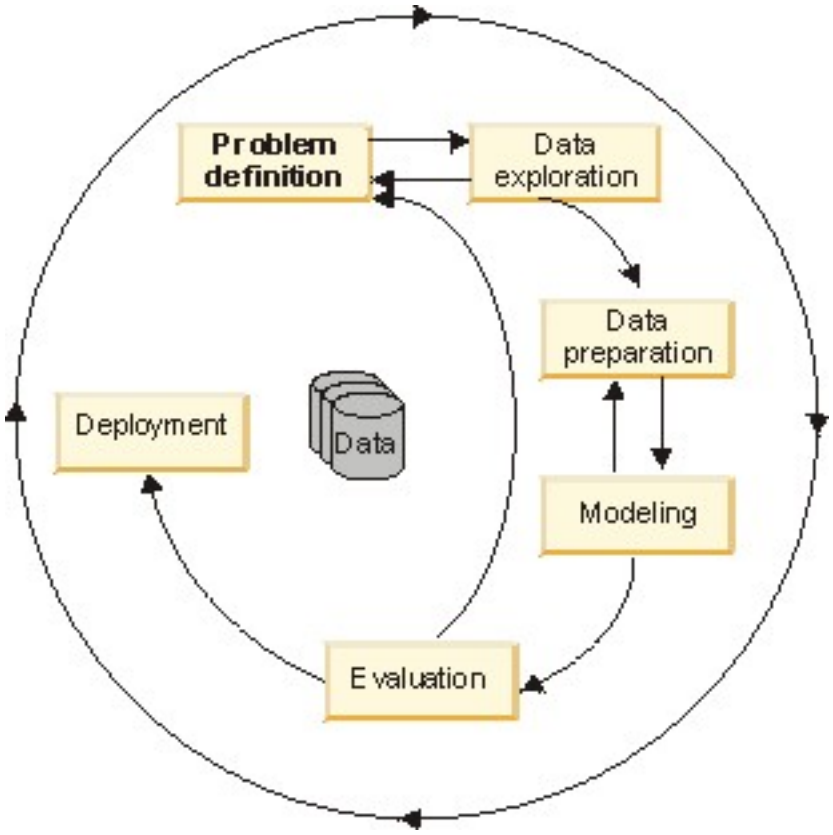
IBM Redbook

Mining Your Own Business in Retail: Using DB2 Intelligent Miner for Data. [ISBN 0738422940]

Chapter 3

The data mining process

To generate new information, you must follow a complete process that runs from a business problem definition to the final deployment of the results that are generated. The following graphic shows this process:



Problem definition A data mining project starts with the understanding of the business problem. Data mining experts, business experts, and domain experts work closely together to define the project objectives and the requirements from a business perspective. The project objective is then translated into a data mining problem definition.

In the problem definition phase, data mining tools are not yet required.

Data exploration Domain experts understand the meaning of the metadata (Glossary on page 63). They collect, describe, and explore the data. They also identify quality problems of the data, for example, missing values or outliers. A frequent exchange with the data mining experts and the business experts from the problem definition phase is vital.

In the data exploration phase, traditional data analysis tools, for example, statistics, are used to identify any quality problems.

Data preparation Domain experts build the data model for the modeling process. They collect, cleanse, and format the data because some of the mining functions accept data only in a certain format. They also create new derived attributes, for example, an average value.

In the data preparation phase, data is tweaked multiple times in no prescribed order. Preparing the data for the modeling tool by selecting tables, records, and attributes, are typical tasks in this phase. The meaning of the data is not changed.

Modeling IT professionals apply various mining techniques to build models because you can use different mining functions for the same type of data mining problem. In our case, we will use the IBM DB2 Easy Mining procedures. Some of the mining functions require specific data types. The data mining experts must assess each model.

In the modeling phase, a frequent exchange with the domain experts from the data preparation phase is required.

The modeling phase and the evaluation phase are coupled. You can repeat these phases several times to change parameters until optimal values are achieved. When the final modeling phase is completed, a model of high quality is build.

Evaluation IT professionals and domain experts with management skills evaluate the model by using visualization tools, the IM Visualization in our case. If the model does not satisfy their expectations, they bo back to the modeling phase and rebuild the model by changing its parameters until optimal values are achieved. When they are finally satisfied with the model, they can extract business explanations and evaluate the following questions:

- Does the model achieve the business objective?
- Have all business issues been considered?

At the end of the evaluation phase, the data mining experts decide how to use the data mining results.

Deployment Depending on the mining function applied, IT professionals whether can apply the model to new data or export the results into database tables or into other applications, for example, spreadsheets. The IM Scoring will help us in this task.

References:

IBM Redbook

Enhance your Business Applications. Simple Integration of Advanced Data Mining Functions. [ISBN 0738427799]

Chapter 1

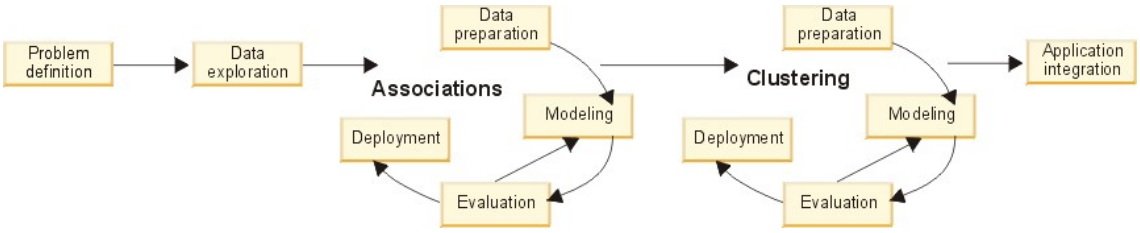
IBM Redbook

Intelligent Miner for Data: Enhance your Business Intelligence. [ISBN 0738413488]

Chapter 2

Tutorial overview

The following graphic shows the structure of this tutorial.



The problem definition phase and the data exploration phase are the same in every data mining project.

In this tutorial two mining functions, among the many different that there are, will be presented. First, you will go through the Associations mining function and then through the Clustering mining function. Both will be applied to the retail company scenario data. You can follow the individual data mining phases (Data preparation, Modeling, Evaluation and Deployment) with each of these mining functions. In our scenario we will only go back from the Evaluation phase to the Modeling if the model built was not good enough for our business problem. Finally, we will finish the process with a new phase, Application integration, which is not in the data mining process but very important to see how to integrate data mining into corporate software solutions.

On the other hand, for the Modeling and Deployment we will use the Easy Mining Procedures. Actually, with them, you can achieve best data mining results easy and fast for almost all business questions without having a deep knowledge of data mining. They help you to create and test mining models (Modeling phase) and then apply these models (Deployment phase) to new data to help you make successful business decisions for the most common business questions in various application areas such as Retail or Manufacturing or Banking.

Section 3. Getting started

Defining the business problem

As already mentioned in the introduction of this tutorial, the scenario we will focus in this case is:

"A retail company wants to have a better knowledge of their clients and their behaviour so that it can offer a better product recommendation to their clients in its next cross selling campaign."

Once the business problem is stated, we need to translate it into a Data Mining problem. In this case, we need to analyze two things: the customer purchasing behaviour and the different customer profiles (and their characteristics). The problem definition can be therefore summarized with these two questions:

"What is the customer purchasing behaviour? Which customers does the company have?"

By answering them, the managers will be able to improve the cross selling campaign they are planning to do. On the one hand, they will be able to offer, for example, the product B when a customer buys product A because the data mining showed that there is a link between the two products. Actually, the product recommendations are based in this linkage, also called association. Therefore, we will apply the ASSOCIATIONS mining technique to find out what do customers purchase together.

On the other hand, knowing better the customers and their general characteristics will help the company to improve its cross-selling campaign (Glossary on page 63), offering special products to special customers. In fact, it is more effective to determine product bundles for groups of customers that have similar behaviours than to try to create a model based on the entire population. To obtain a segmentation of the customers into different groups will be done by applying the CLUSTERING mining technique.

Now that the Data Mining problems have been stated, we can start deciding which Data from the available in our Corporate Database are important to solve the problem.

Data exploration phase: selecting data

In this phase we have to determine which Data (table, views) should be taken into account and collected. If we analyze the questions stated in the former step...

"What is the customer purchasing behaviour? What are the characteristics of the customers?"

... we can deduce that we will need data about the transactions, i.e. customer id, product purchased, quantity, price, etc.

Now we have to identify which tables or views in our database contain all the information we need. This data will be used to create the models in the Modeling phase.

The following diagram, shows the relationship between the tables we have in the Retail scenario.

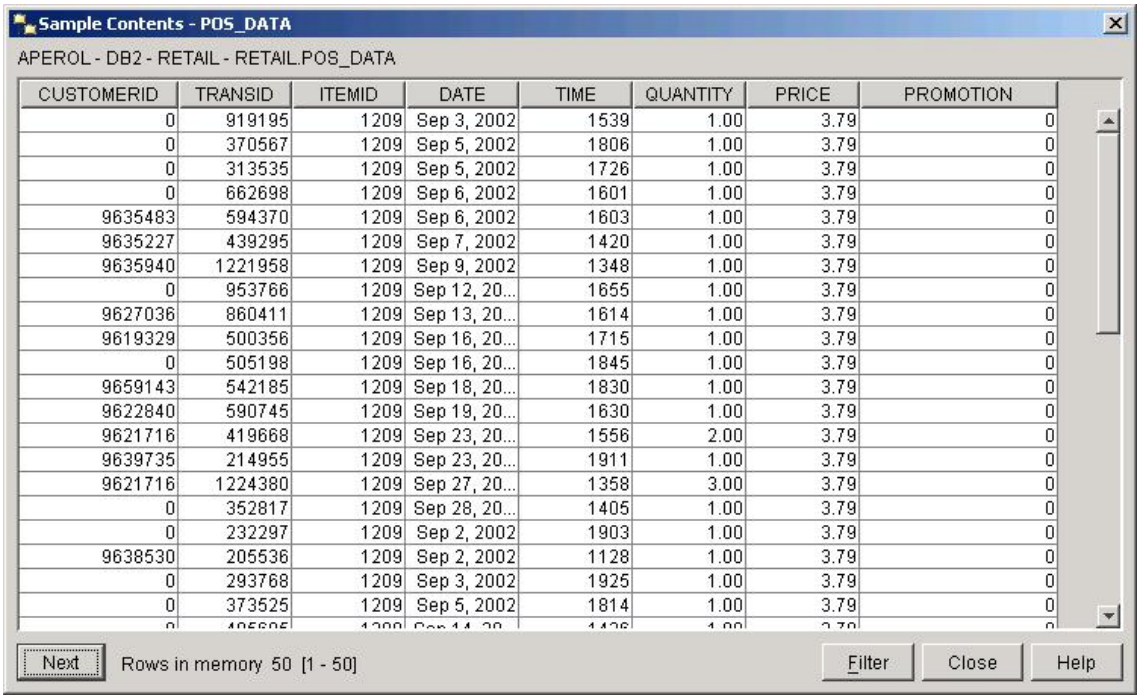
The schema of the tables in the retail database.

POS_DATA contains the information about all the transactions that took place during one month (September 2002). Among this information we find:

RETAIL.POS_DATA

- CUSTOMERID: the customer id.
- TRANSID: the transaction id.
- ITEMID: the item id.
- DATE: the date when the transaction took place.
- TIME: the time when the transaction took place.
- QUANTITY: amount of items with item id = ITEMID, purchased in the transaction.

- PRICE: the price of the item.
- PROMOTION: the discount that was applied to the item.



CUSTOMERID	TRANSID	ITEMID	DATE	TIME	QUANTITY	PRICE	PROMOTION
0	919195	1209	Sep 3, 2002	1539	1.00	3.79	0
0	370567	1209	Sep 5, 2002	1806	1.00	3.79	0
0	313535	1209	Sep 5, 2002	1726	1.00	3.79	0
0	662698	1209	Sep 6, 2002	1601	1.00	3.79	0
9635483	594370	1209	Sep 6, 2002	1603	1.00	3.79	0
9635227	439295	1209	Sep 7, 2002	1420	1.00	3.79	0
9635940	1221958	1209	Sep 9, 2002	1348	1.00	3.79	0
0	953766	1209	Sep 12, 20...	1655	1.00	3.79	0
9627036	860411	1209	Sep 13, 20...	1614	1.00	3.79	0
9619329	500356	1209	Sep 16, 20...	1715	1.00	3.79	0
0	505198	1209	Sep 16, 20...	1845	1.00	3.79	0
9659143	542185	1209	Sep 18, 20...	1830	1.00	3.79	0
9622840	590745	1209	Sep 19, 20...	1630	1.00	3.79	0
9621716	419668	1209	Sep 23, 20...	1556	2.00	3.79	0
9639735	214955	1209	Sep 23, 20...	1911	1.00	3.79	0
9621716	1224380	1209	Sep 27, 20...	1358	3.00	3.79	0
0	352817	1209	Sep 28, 20...	1405	1.00	3.79	0
0	232297	1209	Sep 2, 2002	1903	1.00	3.79	0
9638530	205536	1209	Sep 2, 2002	1128	1.00	3.79	0
0	293768	1209	Sep 3, 2002	1925	1.00	3.79	0
0	373525	1209	Sep 5, 2002	1814	1.00	3.79	0

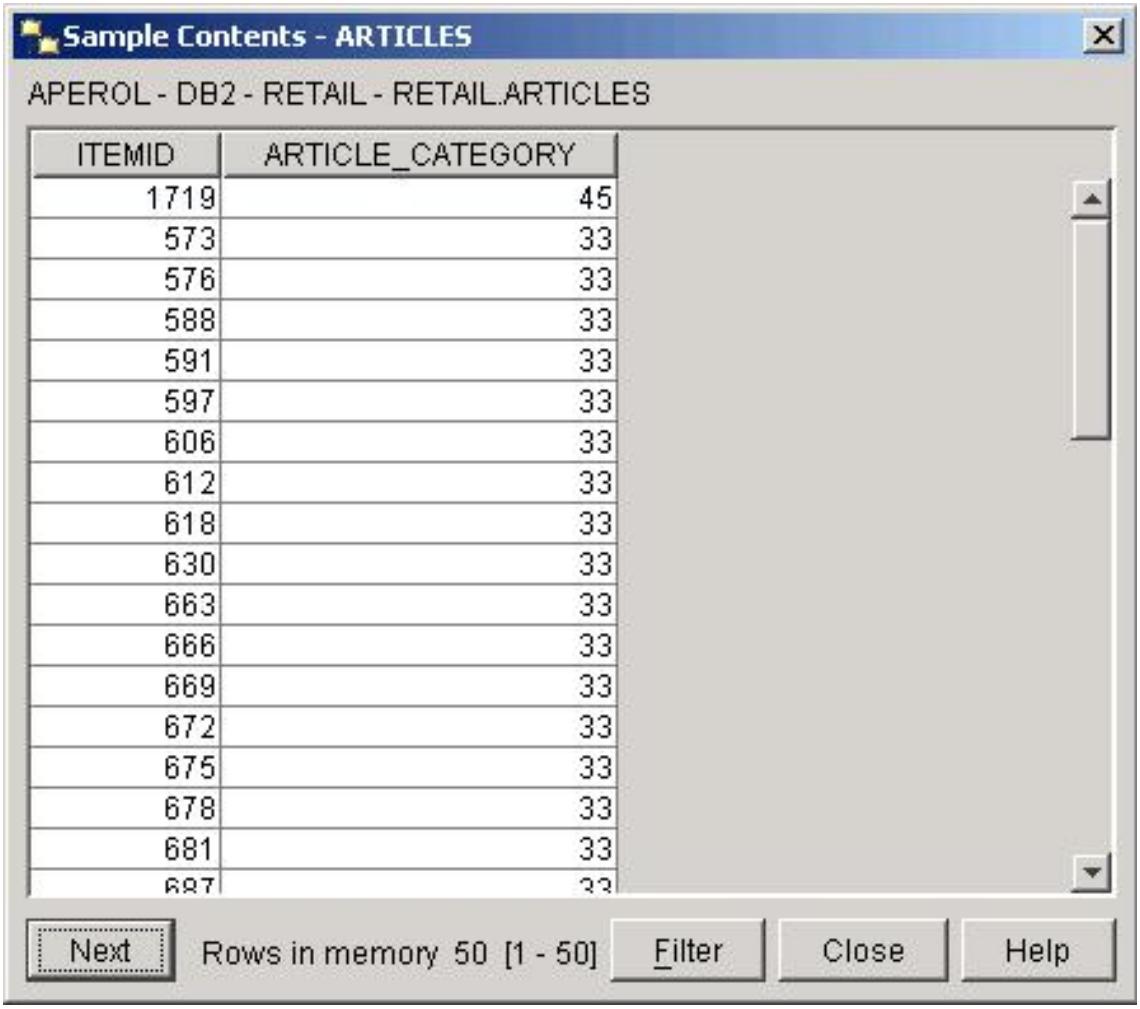
The RETAIL.POS_DATA table.

NOTE: *To see the tables with DB2, go to Start -> Program Files -> IBM DB2-> General Administration Tools -> Control Center, if you are working with Windows. Type db2cc, from a Linux console to start the DB2 Control Center. Select the Retail database in the left frame and go to tables. Select the table from the list and right click your mouse. Select Sample Contents to display it.*

In the table above, the item id is a foreign key to the RETAIL.ARTICLES tables.

RETAIL.ARTICLES

- ITEMID: item id.
- ARTICLE_CATEGORY: an id that points to the ARTICLE_CATEGORIES table, where there is a higher description of the item. For example: milk and beer belong to the beverage category.

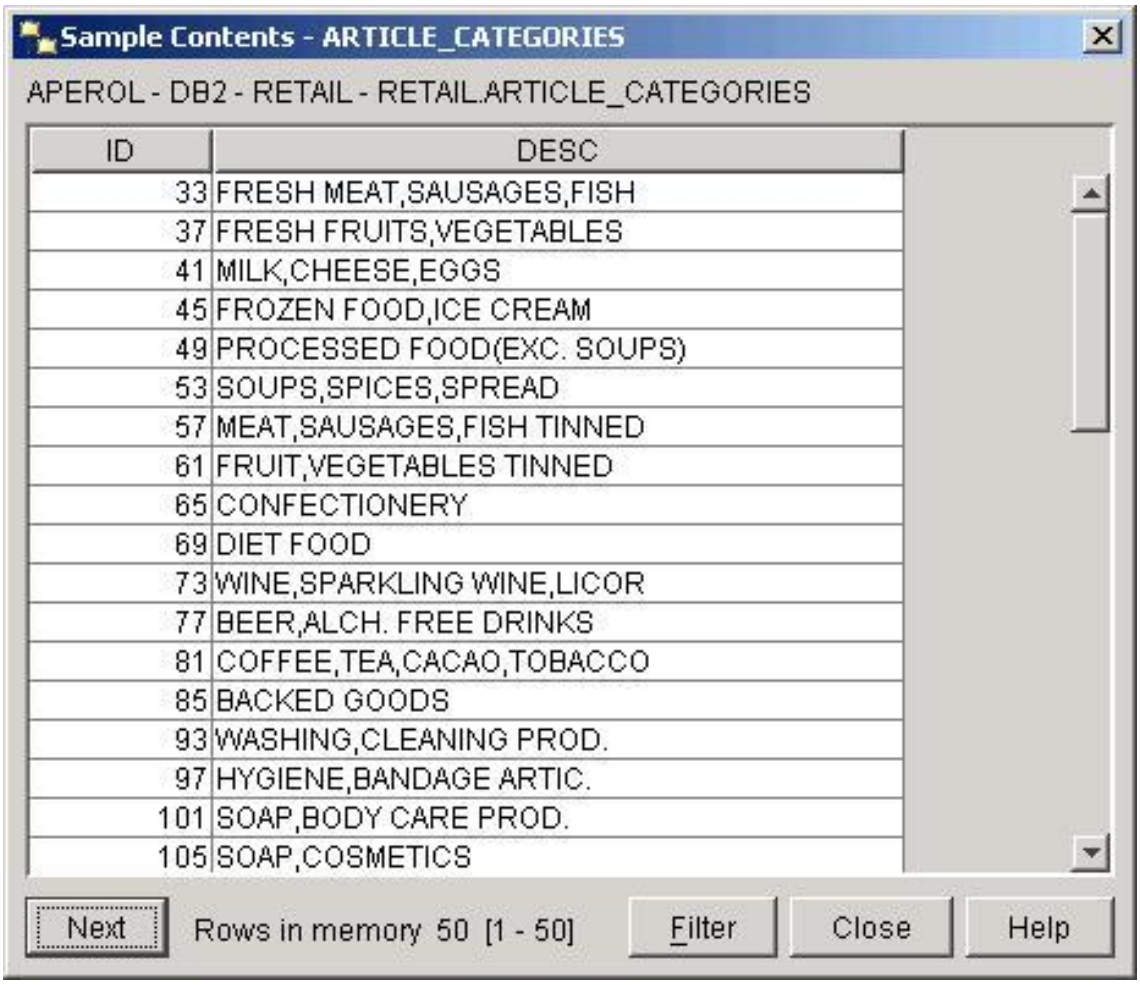


The RETAIL.ARTICLES table.

Finally, the ARTICLES table contain a categoryid attribute that is foreign key to the ARTICLE_CATEGORIES table describe below.

RETAIL.ARTICLE_CATEGORIES

- ID: the id of the article category.
- DESC: the description of the article category.



The RETAIL.ARTICLE_CATEGORIES table.

As indicated in the tutorial overview, the following steps (data preparation, modeling, evaluation and deployment) will be presented twice: once for the ASSOCIATIONS mining function and once for the CLUSTERING mining function.

Section 4. Creating an associations model

Introducing the Associations mining function

The Associations belongs to the Discovery Data Mining techniques, which are used to find patterns inside data without any prior knowledge of which patterns exist. Indeed, Associations or link analysis describes a family of techniques that determine associations between data records.

The most well known type of link analysis is market basket analysis, our scenario. In this case the data records are the items purchased by a customer during the same transaction. The technique discovers the combinations of items that are purchased by different customers. Therefore, by association (or linkage) you can build a picture of which types of product are purchased together. We can obtain rules such as:

RULE

[FROZEN FOOD, ICE-CREAM] + [PET FOOD] ==> [WASHING, CLEANING PROD.].

Support = 0.5371%. Confidence = 38.3475%. Lift = 3.7821

Where:

'+'indicates AND.

'==>'indicates IMPLIES.

Support the percentage of transactions where a customer purchased frozen food or ice-creams, pet food and washing or cleaning products, divided by the total number of transactions and then multiplied by 100 to obtain a percentage value. (see Glossary on page 63 for more)

Confidence the probability that this rule is fulfilled, that is among all the transactions where frozen food or ice-creams and pet food are purchased, how many in percentage also have washing or cleaning products. This could also be expressed as the number of transactions that have frozen food or ice-cream, pet food and washing or cleaning food divided by the number of transactions that have only frozen food or ice-cream and pet food, and then multiplied by 100.(see Glossary on page 63 for more)

Lift a measure of how the rule improves our ability to predict the head of the rule. For example, if the item A and B are purchased together (rule A ==> B) and the Lift is 10, it means that the probability of finding B in those transactions where A is also purchased is 10 times higher than the probability of finding B in all the transactions (no matter if A was purchased or not). Applied it to the previous rule: the probability of finding

washing, cleaning products in those transactions that contain frozen food, ice-cream and pet food is 3.78 times higher the probability of finding it (washing, cleaning products) in any other transaction.(see Glossary on page 63 for more)

NOTE: The item "FROZEN FOOD, ICE-CREAM" refers to the name of the category that contains all the frozen food and ice-creams. The same happens with WASHING, CLEANING PROD.

The former rule could be translated as:

"Customers that purchase frozen food or ice-cream and pet food, buy washing or cleaning products as well in the 38.3475% of the cases. This rule affects the 0.5371% of the transactions we are analyzing. Furthermore, the probability of finding washing or cleaning products in the transactions where frozen food or ice-cream and pet food was purchased, is 3.7821 times higher as the probability of finding it in the any other transaction."

With the Support, Confidence and Lift parameters, you will be able to decide which rules are more important for your business problem. As a guideline, we will present a general priority list of parameters to determine whether a rule is interesting or not. However, this list will depend on the business problem we are facing.

In the next section, Data Preparation, we will see how we select and prepare the data before we do modeling.

Data preparation phase

This step consists in the collection of the data (tables, views) which will help us to solve the problem. Moreover, the checking of possible invalid values (outliers, missing values, etc.) that we did in the last step (Data Exploration) will help us now to clean and format the data. We can also generate new aggregated attributes such as averages or subtractions, for example.

In our scenario, we will take into account all the transactions, despite some are done by customers without a customer card. This is not a problem, since we want to analyze customer behaviour, we focus in the items that are purchased together in each transaction.

For the associations, furthermore, the input table or view for the model **must contain only two columns**, the first will be the transaction id and the second an item that contains the elements that we want to use to obtain rules. In our case this will be the

category id of each product.

The RETAIL.POS_DATA contains almost all the information we need: the transaction id and the item id. This last attribute (item id) will have to be translated into category id, using the RETAIL.ARTICLES table.

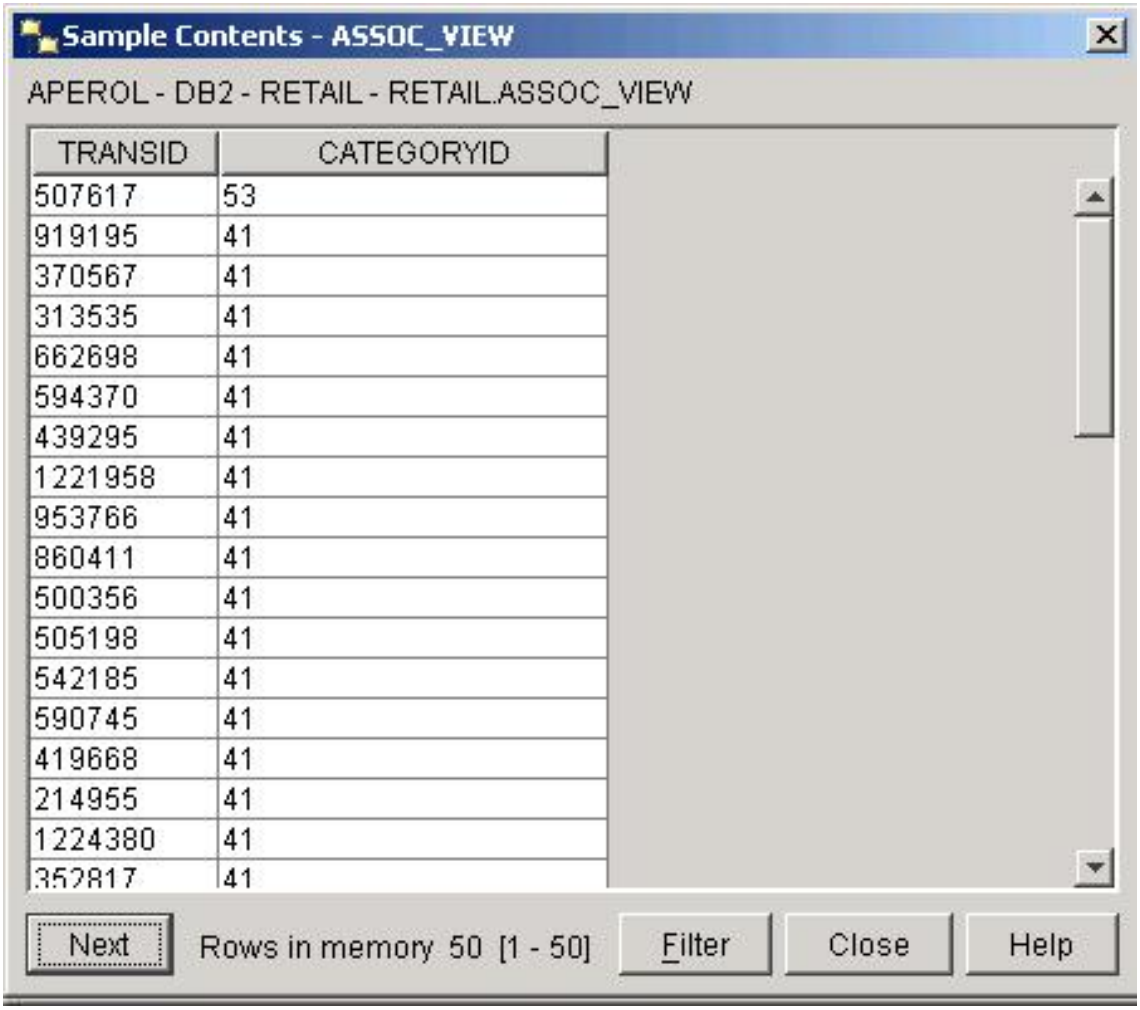
So, we need a table or view that has for each transaction, the transaction id and the category id's of the items that were purchased, instead of the item id's. To obtain such a view or table, execute the following command:

```
[tutorial_path]:\> db2 -tvf retail_assoc_view.db2
```

You can click Troubleshooting on page 62 to check for errors that might occur when you use the above SQL command.

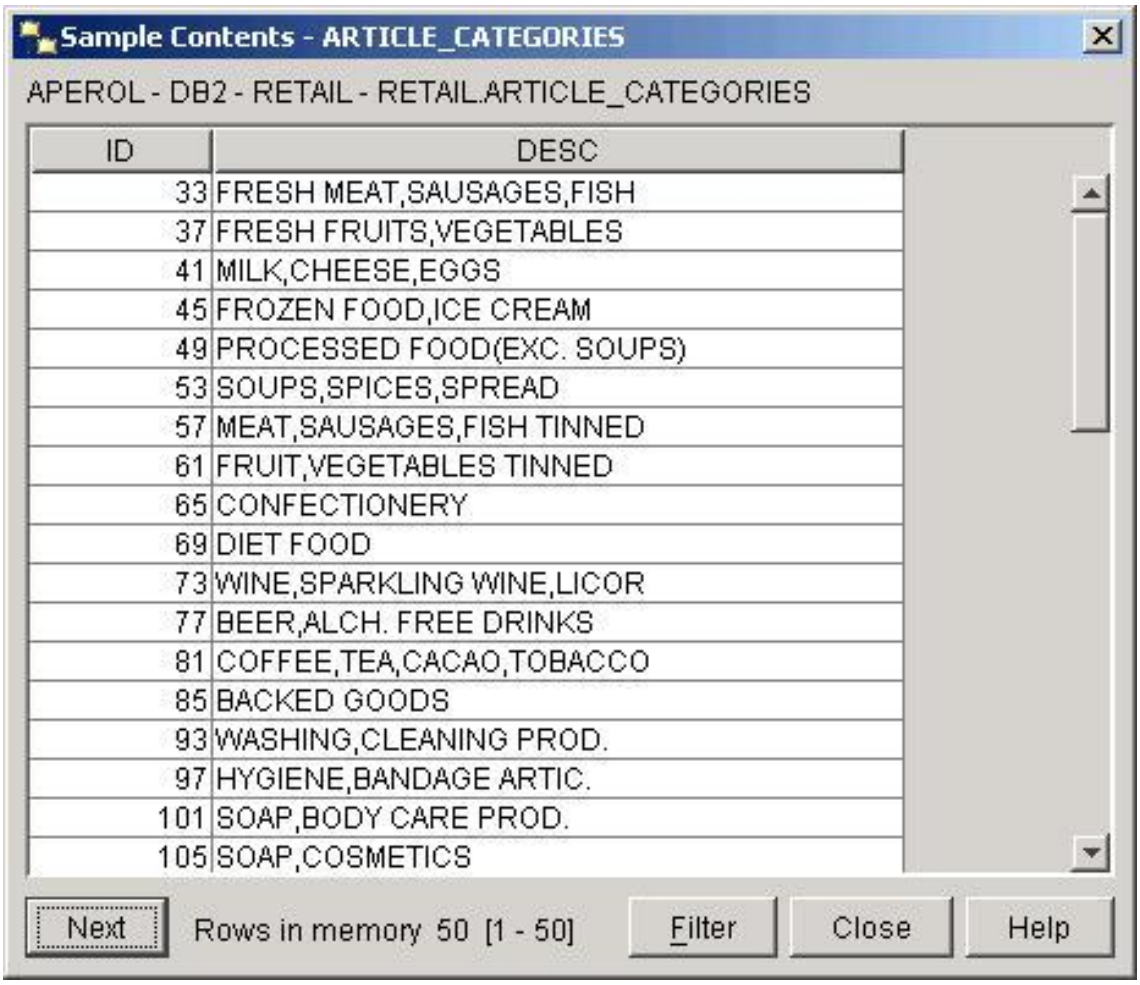
To view the content of this script, click *here*.

As a result of the execution we will obtain the following view:



RETAIL.ASSOC_VIEW.

Yet this is not enough. If we applied ASSOCIATIONS now, we would obtain rules only with the categories ids. The desirable result should include a description of each category, so that the rules are more legible. We will do this, using a Name Mapping when building the model in the next step. The mapping will be done between the category id that appears in the view above and the category description that is in the RETAIL.ARTICLE_CATEGORIES table. Here is a screenshot of this last table:



RETAIL.ARTICLE_CATEGORIES.

Modeling phase: using Easy Mining procedures

The Easy Mining Procedure to build an ASSOCIATIONS model

When the Mining technique is chosen, we have to build the model. In this section will present how easy it is to build ASSOCIATIONS models with the Easy Mining Procedures of the IBM DB2 Intelligent Miner.

There are many procedures to build associations models, which can be found in the *RuleProcsDecl.sql file*. We will focus on the following procedure call:

```
IDMMX.BuildRuleModel(<modelName>, <tableName>, <groupColumn>,
<minSupport>, <minConfidence>, <maxRuleLength>, <optionalParameters>)
```

Where:

<modelName> is the name that we will give to the model built.

<tableName> is the table or view that contains the Data Model (RETAIL_ASSOC_VIEW for the associations in the retail case).

<groupColumn> is the primary key of the records in the table (the TRANS_ID attribute).

<minSupport> specifies the minimum support percentage that the rules must have, that is how frequently the rule occurs among all of the groups or which percentage of the population is affected by that rule. Notice that this function takes a percentage value, hence minSupport=25 will create rules with a minimum support of 25%. The given value must have a DOUBLE type.

<minConfidence> sets the minimum confidence that the rules generated must have, that is, the likelihood that the rule is fulfilled. Here again, the function takes a percentage value. The value must be DOUBLE.

<maxRuleLength> sets the maximum number or items that a rule can have, including head and body, (see Glossary on page 63 for more). The value must be INTEGER.

<optionalParameters> among others, can be:

The output of this procedure call will be an ASSOCIATIONS model that will be stored in an output table, the IDMMX.RULEMODELS table.

Modeling phase: building an associations model

Building an associations model for our Scenario

In our retail scenario, we can call:

```
[tutorial path]:\> db2 -tvf buildrulemodel.db2
```

You can click Troubleshooting on page 62 to check for errors that might occur when you use the above SQL command.

The content of the script is:

```
connect to RETAIL;
call IDMMX.BuildRuleModel('RETAIL.ASSOC_MODEL','RETAIL.ASSOC_VIEW',
'TRANSID', 0.5, 25, 3,
'DM_addNmp(''NewMap'', ''RETAIL.ARTICLE_CATEGORIES'', ''ID'', ''DESC''),
DM_setFldNmp(''CATEGORYID'', ''NewMap'')');
connect reset;
```

Note: '' are two single quotes.

This call will generate a model with rules, each of which will have at most 3 items (including head and body), a minimum confidence of the 25% and will be supported by a minimum of the 0.1% of the transactions in the *RETAIL.ASSOC_VIEW*.

We have set the Confidence to a 25% and the Support to a 0.5%, as a first approach. The higher we set the confidence and the support constraints, the less rules we would obtain. Probably the model will generate a lot of rules with the values given, we will see it in the Evaluation step. If we consider the model not appropriate (the model provides too many rules, irrelevant rules for the business problem, etc.), we can rebuild it with a higher support and/or confidence.

Furthermore, we have defined a name mapping. In this case, we will translate the item id into it's category description. This means, that instead of obtaining rules with the item id of each product, we will obtain rules with the name of the category that the product belongs to, as we said in the Data Preparation.

At the end of the Data Preparation and modeling phases, a model has been built and stored in the IDMMX.RULEMODELS table of our database. In the next step (Evaluation) we will assess the model with a graphical report that the IBM DB2 Visualization generates.

Evaluation phase: using IM Visualization

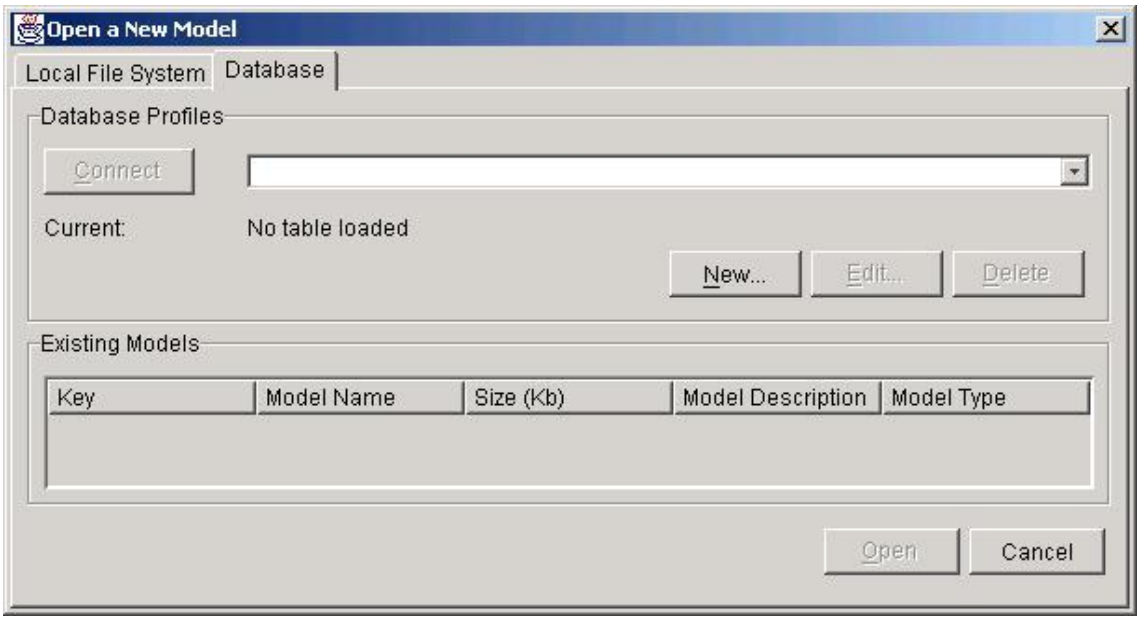
When the model is created, you can use IBM DB2 Intelligent Miner Visualization to look at the results and to evaluate whether the model is good. To start the IM Visualizer, go to **Program Files -> IBM DB2 Intelligent Miner Visualization V8.1 -> IM Visualizer** if you are working with Windows, or execute **imvisualizer** from a console if you are working with Linux.

Connecting to a database

Note: Before you start the IM visualization, ensure that you have copied the following files into the *\Program Files\IBM\IMVisualization\lib* directory for Windows, and in the the */IMVisualization/lib* if you are using Linux. These files are stored in the *\Program Files\IBM\SQLLIB\java* directory (Windows) or in the */IBM/db2/V8.1/java* directory (Linux).

- db2java.zip
- db2jcc.jar

When you start the IM Visualization, click the **Database** label of the window displayed:



The Open a New Model window

On the Open a New Model window, you can select one of the models in the *Existing models* list. If you use the IM Visualizer for the first time, the Existing Models list is empty. Before you can view the existing models, you must create database profiles.

To create a new database profile, click **New...**.

Creating a new database profile

On the Add New Database Profile window, type the following parameters in the appropriate entry fields and click **Retrieve Database Information**:

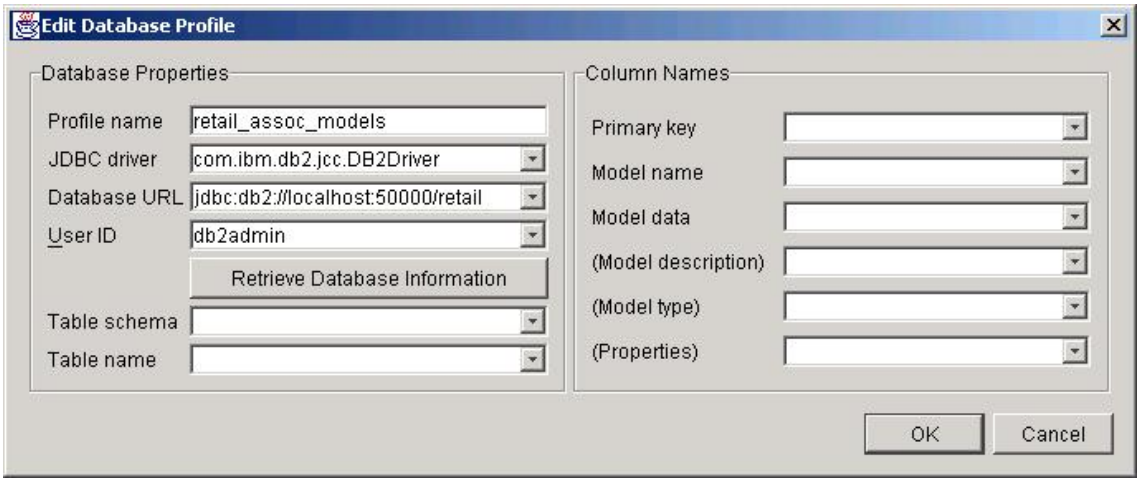
Note: The name of the JDBC driver is automatically displayed, but make sure that corresponds to the one specified below. If you cannot create the Database Profile, it might be that your JDBC driver and/or you URL are different. In this case, contact your system administrator.

Profile name retail_assoc_models

JDBC driver com.ibm.db2.jcc.DB2Driver

Database URL jdbc:db2://localhost:50000/retail (localhost, if DB2 is installed localy in your machine. If not, contact your system adminnistrator)

UserID db2admin (for example)



The Add New Database Profile window

Type the following parameters in the appropriate entry fields and click **OK** to return to the Open a New Model window:

Table schema IDMMX

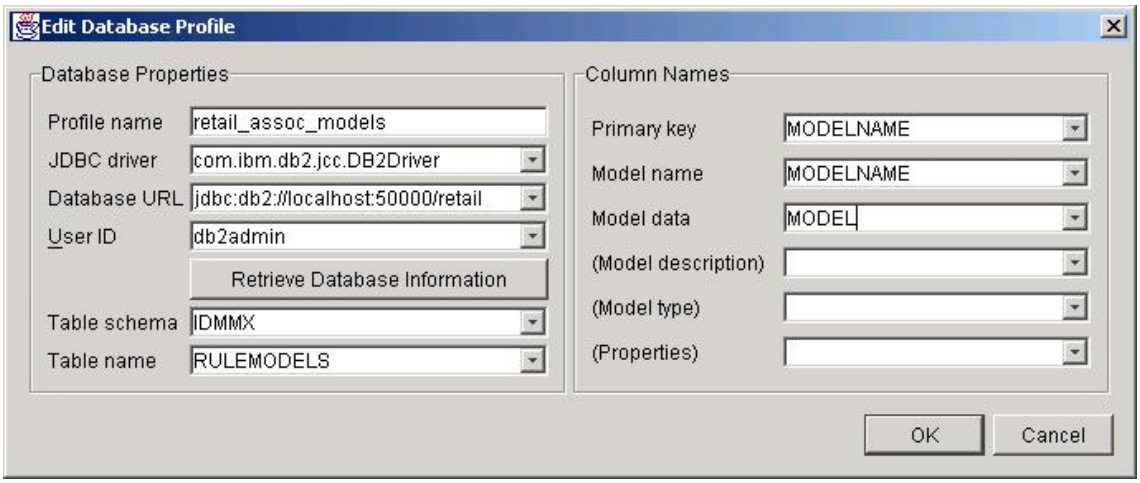
Table name RULEMODELS

Primary key MODELNAME

Model name MODELNAME

Model data MODEL

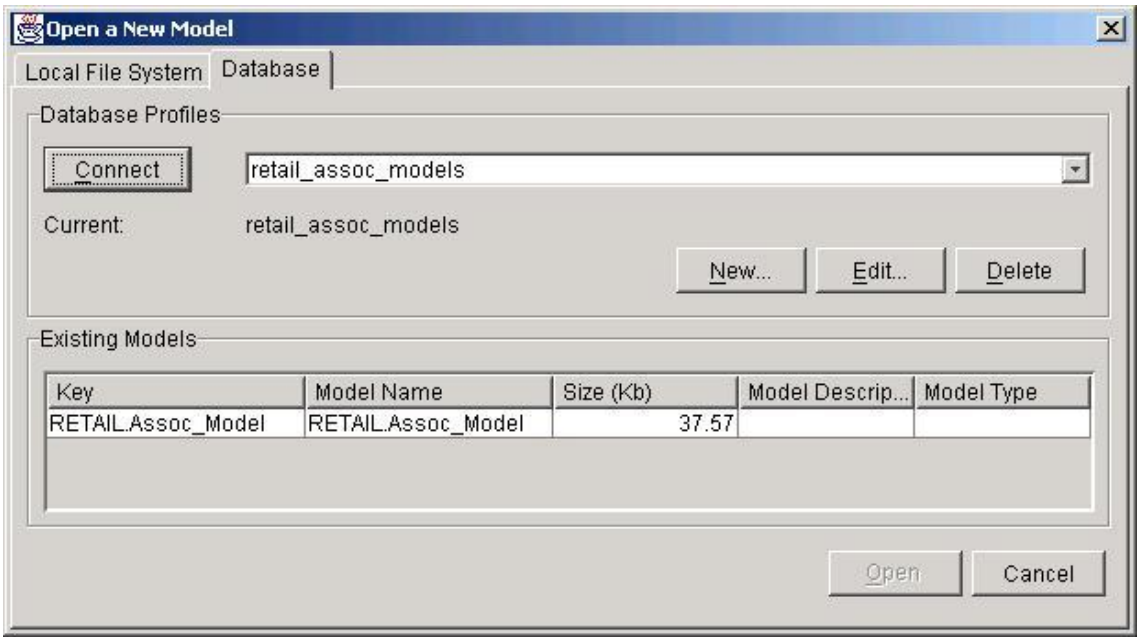
The following graphic shows the Add New Database Profile window with the complete specification:



The complete specification of the Add New Database Profile window

Opening a model

On the Add New Database Profile window, click **Connect** to connect to the database that you have specified on the previous window. When your operating system is connected to the database, a list of models that are available in this database is displayed.



The Open a New Model window displaying a list of existing models

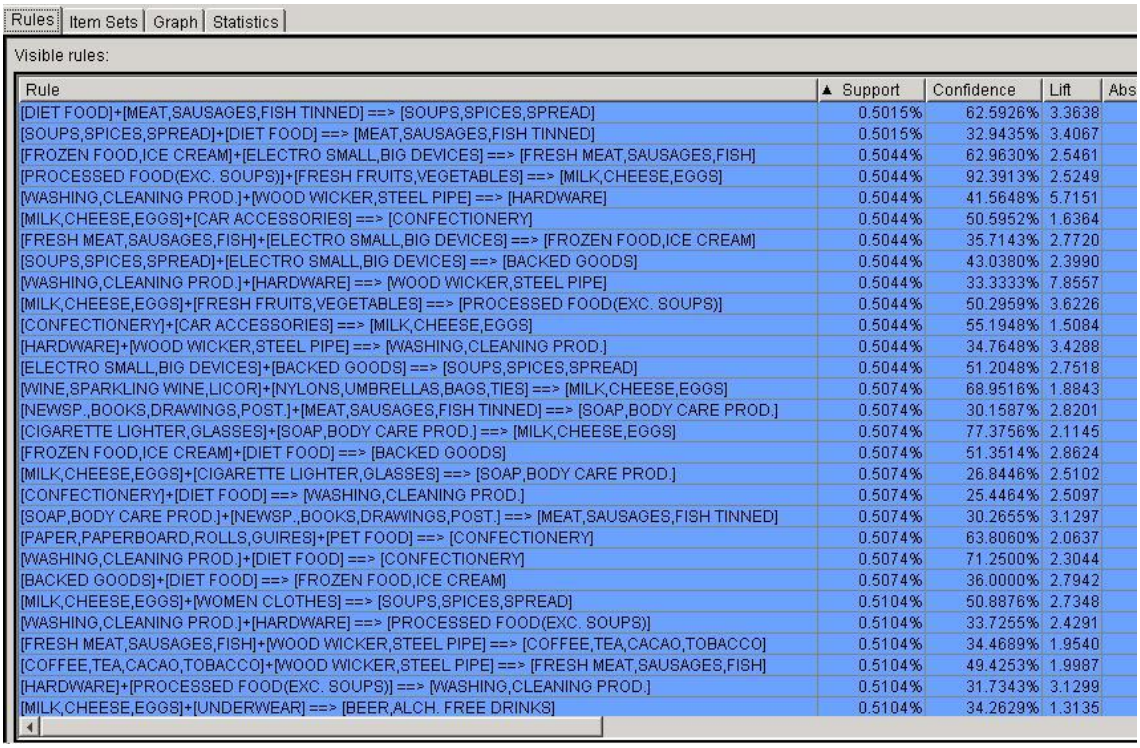
To open the RETAIL.ASSOC_MODEL, select this model from the list of existing models and click **Open**.

Evaluation phase: evaluating the results

The evaluation of the model consists in the evaluation of the rules the model has generated as a result. Our task now is to choose which ones provide with an interesting knowledge for our business problem. That is why in this step you will have to analyze the results alongside with the business experts.

Graphical Report of the Results:

Let's see the results of the model in a graphical report the IM Visualization generates.



Rules | Item Sets | Graph | Statistics

Visible rules:

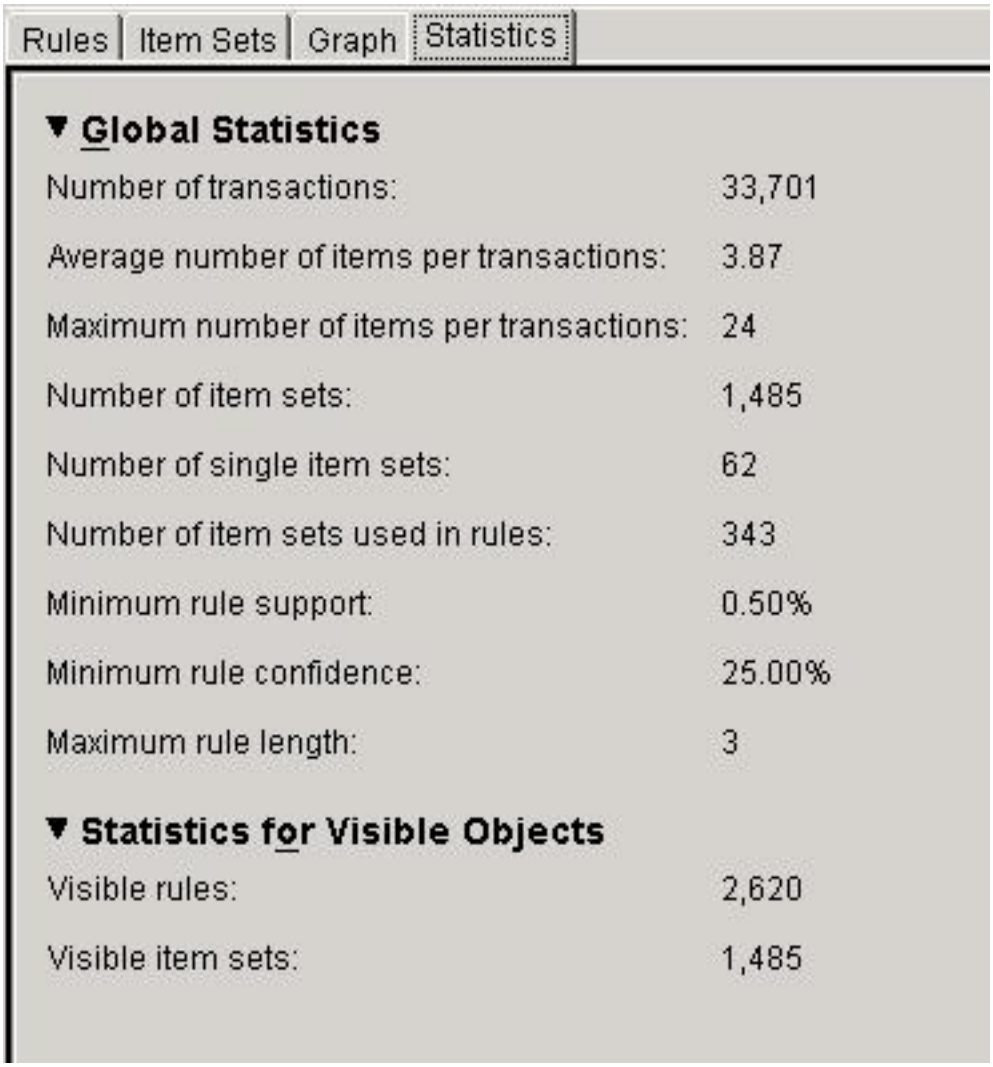
Rule	Support	Confidence	Lift	Abs
[DIET FOOD]+[MEAT,SAUSAGES,FISH TINNED] ==> [SOUPS,SPICES,SPREAD]	0.5015%	62.5926%	3.3638	
[SOUPS,SPICES,SPREAD]+[DIET FOOD] ==> [MEAT,SAUSAGES,FISH TINNED]	0.5015%	32.9435%	3.4067	
[FROZEN FOOD,ICE CREAM]+[ELECTRO SMALL,BIG DEVICES] ==> [FRESH MEAT,SAUSAGES,FISH]	0.5044%	62.9630%	2.5461	
[PROCESSED FOOD(EXC. SOUPS)]+[FRESH FRUITS,VEGETABLES] ==> [MILK,CHEESE,EGGS]	0.5044%	92.3913%	2.5249	
[WASHING,CLEANING PROD.]+[WOOD WICKER,STEEL PIPE] ==> [HARDWARE]	0.5044%	41.5648%	5.7151	
[MILK,CHEESE,EGGS]+[CAR ACCESSORIES] ==> [CONFECTIONERY]	0.5044%	50.5952%	1.6364	
[FRESH MEAT,SAUSAGES,FISH]+[ELECTRO SMALL,BIG DEVICES] ==> [FROZEN FOOD,ICE CREAM]	0.5044%	35.7143%	2.7720	
[SOUPS,SPICES,SPREAD]+[ELECTRO SMALL,BIG DEVICES] ==> [BACKED GOODS]	0.5044%	43.0380%	2.3990	
[WASHING,CLEANING PROD.]+[HARDWARE] ==> [WOOD WICKER,STEEL PIPE]	0.5044%	33.3333%	7.8557	
[MILK,CHEESE,EGGS]+[FRESH FRUITS,VEGETABLES] ==> [PROCESSED FOOD(EXC. SOUPS)]	0.5044%	50.2959%	3.6226	
[CONFECTIONERY]+[CAR ACCESSORIES] ==> [MILK,CHEESE,EGGS]	0.5044%	55.1948%	1.5084	
[HARDWARE]+[WOOD WICKER,STEEL PIPE] ==> [WASHING,CLEANING PROD.]	0.5044%	34.7648%	3.4288	
[ELECTRO SMALL,BIG DEVICES]+[BACKED GOODS] ==> [SOUPS,SPICES,SPREAD]	0.5044%	51.2048%	2.7518	
[WINE,SPARKLING WINE,LICOR]+[NYLONS,UMBRELLAS,BAGS,TIES] ==> [MILK,CHEESE,EGGS]	0.5074%	68.9516%	1.8843	
[NEWSP.,BOOKS,DRAWINGS,POST.]+[MEAT,SAUSAGES,FISH TINNED] ==> [SOAP,BODY CARE PROD.]	0.5074%	30.1587%	2.8201	
[CIGARETTE LIGHTER,GLASSES]+[SOAP,BODY CARE PROD.] ==> [MILK,CHEESE,EGGS]	0.5074%	77.3756%	2.1145	
[FROZEN FOOD,ICE CREAM]+[DIET FOOD] ==> [BACKED GOODS]	0.5074%	51.3514%	2.8624	
[MILK,CHEESE,EGGS]+[CIGARETTE LIGHTER,GLASSES] ==> [SOAP,BODY CARE PROD.]	0.5074%	26.8446%	2.5102	
[CONFECTIONERY]+[DIET FOOD] ==> [WASHING,CLEANING PROD.]	0.5074%	25.4464%	2.5097	
[SOAP,BODY CARE PROD.]+[NEWSP.,BOOKS,DRAWINGS,POST.] ==> [MEAT,SAUSAGES,FISH TINNED]	0.5074%	30.2655%	3.1297	
[PAPER,PAPERBOARD,ROLLS,GUIRES]+[PET FOOD] ==> [CONFECTIONERY]	0.5074%	63.8060%	2.0637	
[WASHING,CLEANING PROD.]+[DIET FOOD] ==> [CONFECTIONERY]	0.5074%	71.2500%	2.3044	
[BACKED GOODS]+[DIET FOOD] ==> [FROZEN FOOD,ICE CREAM]	0.5074%	36.0000%	2.7942	
[MILK,CHEESE,EGGS]+[WOMEN CLOTHES] ==> [SOUPS,SPICES,SPREAD]	0.5104%	50.8876%	2.7348	
[WASHING,CLEANING PROD.]+[HARDWARE] ==> [PROCESSED FOOD(EXC. SOUPS)]	0.5104%	33.7255%	2.4291	
[FRESH MEAT,SAUSAGES,FISH]+[WOOD WICKER,STEEL PIPE] ==> [COFFEE,TEA,CACAO,TOBACCO]	0.5104%	34.4689%	1.9540	
[COFFEE,TEA,CACAO,TOBACCO]+[WOOD WICKER,STEEL PIPE] ==> [FRESH MEAT,SAUSAGES,FISH]	0.5104%	49.4253%	1.9987	
[HARDWARE]+[PROCESSED FOOD(EXC. SOUPS)] ==> [WASHING,CLEANING PROD.]	0.5104%	31.7343%	3.1299	
[MILK,CHEESE,EGGS]+[UNDERWEAR] ==> [BEER,ALCH. FREE DRINKS]	0.5104%	34.2629%	1.3135	

The rules generated by the RETAIL.ASSOC_MODEL

Let's have a look at the statistics of the model to know how many rules were generated.

Statistics of the model:

If we click at the Statistics label we see under the "Statistics for Visible Objects" that 2,620 rules were generated. If we want to obtain more rules we only have to decrease the Confidence and/or Support constraints when building the model.



The statistics of the RETAIL.ASSOC_MODEL.

Evaluation phase: selecting interesting rules

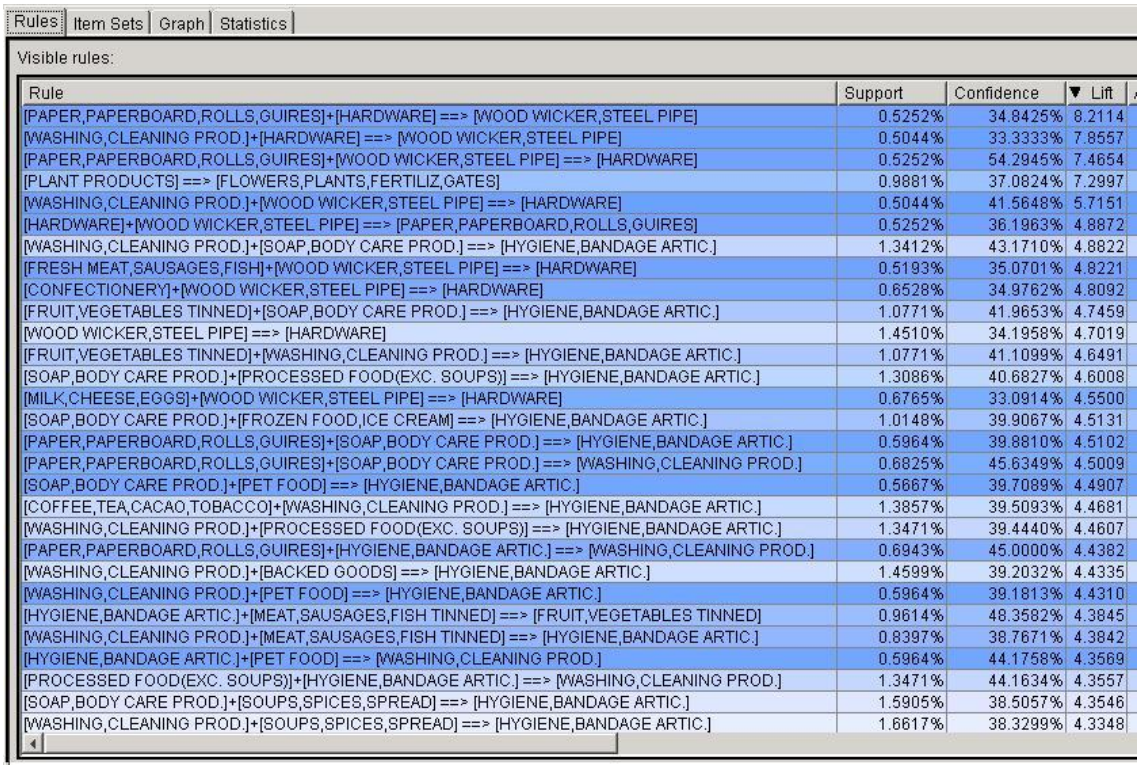
It is clear that we want to select among all the rules generated, those with the highest business value. The IM Visualization allows you to sort the rules according to the *Lift* or the *Support* or the *Confidence* values just clicking these column names.

As a general rule, the higher the lift, the more interesting a rule becomes. In fact, we will consider interesting rules, those with:

1. **A higher Lift**
2. **A higher Support**
3. **A higher Confidence**

NOTE: Depending on the business problem, you may consider that the Confidence is more important than the Support. In our retail scenario, those rules that affect higher transactions (higher support) are more interesting than those that affect less. Obviously, we will not give more priority to those rules that are fulfilled with more confidence but do not affect a lot of customers.

If you sort the rules by their lift, support and confidence, by clicking at these column names, you obtain:



Rules | Item Sets | Graph | Statistics

Visible rules:

Rule	Support	Confidence	▼ Lift
[PAPER,PAPERBOARD,ROLLS,GUIRES]+[HARDWARE] ==> [WOOD WICKER,STEEL PIPE]	0.5252%	34.8425%	8.2114
[WASHING,CLEANING PROD.]+[HARDWARE] ==> [WOOD WICKER,STEEL PIPE]	0.5044%	33.3333%	7.8557
[PAPER,PAPERBOARD,ROLLS,GUIRES]+[WOOD WICKER,STEEL PIPE] ==> [HARDWARE]	0.5252%	54.2945%	7.4654
[PLANT PRODUCTS] ==> [FLOWERS,PLANTS,FERTILIZ,GATES]	0.9881%	37.0824%	7.2997
[WASHING,CLEANING PROD.]+[WOOD WICKER,STEEL PIPE] ==> [HARDWARE]	0.5044%	41.5648%	5.7151
[HARDWARE]+[WOOD WICKER,STEEL PIPE] ==> [PAPER,PAPERBOARD,ROLLS,GUIRES]	0.5252%	36.1963%	4.8872
[WASHING,CLEANING PROD.]+[SOAP,BODY CARE PROD.] ==> [HYGIENE,BANDAGE ARTIC.]	1.3412%	43.1710%	4.8822
[FRESH MEAT,SAUSAGES,FISH]+[WOOD WICKER,STEEL PIPE] ==> [HARDWARE]	0.5193%	35.0701%	4.8221
[CONFECTIONERY]+[WOOD WICKER,STEEL PIPE] ==> [HARDWARE]	0.6528%	34.9762%	4.8092
[FRUIT,VEGETABLES TINNED]+[SOAP,BODY CARE PROD.] ==> [HYGIENE,BANDAGE ARTIC.]	1.0771%	41.9653%	4.7459
[WOOD WICKER,STEEL PIPE] ==> [HARDWARE]	1.4510%	34.1958%	4.7019
[FRUIT,VEGETABLES TINNED]+[WASHING,CLEANING PROD.] ==> [HYGIENE,BANDAGE ARTIC.]	1.0771%	41.1099%	4.6491
[SOAP,BODY CARE PROD.]+[PROCESSED FOOD(EXC. SOUPS)] ==> [HYGIENE,BANDAGE ARTIC.]	1.3086%	40.6827%	4.6008
[MILK,CHEESE,EGGS]+[WOOD WICKER,STEEL PIPE] ==> [HARDWARE]	0.6765%	33.0914%	4.5500
[SOAP,BODY CARE PROD.]+[FROZEN FOOD,ICE CREAM] ==> [HYGIENE,BANDAGE ARTIC.]	1.0148%	39.9067%	4.5131
[PAPER,PAPERBOARD,ROLLS,GUIRES]+[SOAP,BODY CARE PROD.] ==> [HYGIENE,BANDAGE ARTIC.]	0.5964%	39.8810%	4.5102
[PAPER,PAPERBOARD,ROLLS,GUIRES]+[SOAP,BODY CARE PROD.] ==> [WASHING,CLEANING PROD.]	0.6825%	45.6349%	4.5009
[SOAP,BODY CARE PROD.]+[PET FOOD] ==> [HYGIENE,BANDAGE ARTIC.]	0.5667%	39.7089%	4.4907
[COFFEE,TEA,CACAO,TOBACCO]+[WASHING,CLEANING PROD.] ==> [HYGIENE,BANDAGE ARTIC.]	1.3857%	39.5093%	4.4681
[WASHING,CLEANING PROD.]+[PROCESSED FOOD(EXC. SOUPS)] ==> [HYGIENE,BANDAGE ARTIC.]	1.3471%	39.4440%	4.4607
[PAPER,PAPERBOARD,ROLLS,GUIRES]+[HYGIENE,BANDAGE ARTIC.] ==> [WASHING,CLEANING PROD.]	0.6943%	45.0000%	4.4382
[WASHING,CLEANING PROD.]+[BACKED GOODS] ==> [HYGIENE,BANDAGE ARTIC.]	1.4599%	39.2032%	4.4335
[WASHING,CLEANING PROD.]+[PET FOOD] ==> [HYGIENE,BANDAGE ARTIC.]	0.5964%	39.1813%	4.4310
[HYGIENE,BANDAGE ARTIC.]+[MEAT,SAUSAGES,FISH TINNED] ==> [FRUIT,VEGETABLES TINNED]	0.9614%	48.3582%	4.3845
[WASHING,CLEANING PROD.]+[MEAT,SAUSAGES,FISH TINNED] ==> [HYGIENE,BANDAGE ARTIC.]	0.8397%	38.7671%	4.3842
[HYGIENE,BANDAGE ARTIC.]+[PET FOOD] ==> [WASHING,CLEANING PROD.]	0.5964%	44.1758%	4.3569
[PROCESSED FOOD(EXC. SOUPS)]+[HYGIENE,BANDAGE ARTIC.] ==> [WASHING,CLEANING PROD.]	1.3471%	44.1634%	4.3557
[SOAP,BODY CARE PROD.]+[SOUPS,SPICES,SPREAD] ==> [HYGIENE,BANDAGE ARTIC.]	1.5905%	38.5057%	4.3546
[WASHING,CLEANING PROD.]+[SOUPS,SPICES,SPREAD] ==> [HYGIENE,BANDAGE ARTIC.]	1.6617%	38.3299%	4.3348

The rules generated by the RETAIL.ASSOC_MODEL

Let's analyze two rules we have obtained. The first may be seen as evident, whereas the second might give us unknown knowledge:

Example 1

[PLANT PRODUCTS] == > [FLOWERS,PLANTS,FERTILIZ,GATES]

Support = 0.9881% Confidence = 37.0824% Lift = 7.2997 Meaning

Customers that buy plant products purchase flowers, plants, fertilizers or gates as well in the 37.0824% of the cases. This rule affects the 0.9881% of the transactions and the probability to find flowers or plants or fertilizers or gates in those transactions with plant products is 7,2997 times higher as the probability to find them (plant products) in any other transaction. This rule is logical because both products are highly related.

Example 2

[BACKED GOODS] + [PET FOOD] ==> [HYGIENE, BANDAGE ARTIC.]

Support = 0.6854% Confidence = 33.0472% Lift = 3.7373 Meaning

Customers that buy backed goods and pet food purchase hygiene or bandage articles as well in the 33.0472% of the cases. This affects the 0.6854% of the total transactions. Moreover, the probability to hygiene or bandage articles in those transactions with backed goods and pet food is 3.7373 times higher than the probability to find them (hygiene or bandage articles) in any other transactions.

In the next step, we will generate product recommendations for some transactions.

Deployment phase: how to extract the rules into a table

As deployment we will imagine the following situation:

A new customer goes into the physical supermarket and whenever he/she adds products into the cart, a real time application generates product recommendations applying the rules that the associations model built.

To do this, we need to extract the rules into a database table and then scan the table looking for rules that have in the body (or antecedent) one item that is in the cart. Then, we will add the head of the rule to the product recommendation list.

We can extract the associations model that the IBM DB2 modeling has generated in a PMML format (Glossary on page 63) and stored in the IDMMX.RULEMODELS table into a DB2 database table. In the setup of the database we enabled it for the XML-Extender from DB2. Using the functions in this extender we can extract the information from the PMML-Models.

To get the information from the model and transfer it into a database table, execute:

```
[tutorial path]:\> db2 -tvf extract_rules.db2
```

You can click Troubleshooting on page 62 to check for errors that might occur when you use the above SQL command.

To view the content of this script, click *here*.

With this execution we obtain a table with the rules of our model:

Sample Contents - ORDEREDRULES

APEROL - DB2 - RETAIL - RETAIL.ORDEREDRULES

ID	RULEBODY1	RULEBODY2	RULEHEAD	SUP	CONF	LIFT
762	41	101	77	2.819	42.735	1.638
199		81	77	6.222	35.273	1.352
200	81	205	77	0.914	39.896	1.529
19	77	205	81	0.914	38.987	2.21
214	61	81	77	1.777	46.579	1.786
33	61	77	81	1.777	42.035	2.383
1	77	81	61	1.777	28.565	2.59
229	73	81	77	1.406	44.675	1.713
48	73	77	81	1.406	33.617	1.906
243	41	81	77	4.308	40.695	1.56
61	41	77	81	4.308	34.149	1.936
1	77	81	41	4.308	69.242	1.892
264	33	81	77	3.19	43.052	1.65
82	33	77	81	3.19	36.465	2.067
1	77	81	33	3.19	51.264	2.073
282	81	93	77	1.466	41.794	1.602
100	77	93	81	1.466	43.295	2.454
294	81	285	77	0.516	31.127	1.193
111	77	285	81	0.516	29.542	1.675
302	81	101	77	1.579	43.714	1.676
118	77	101	81	1.579	41.89	2.375

Next Rows in memory 50 [1 - 50] Filter Close Help

RETAIL.OrderedRules table: contains the rules generated by the RETAIL.ASSOC_MODEL2.

Support and Confidence are written in % values.

Deployment phase: how to apply the rules to some transactions

Before applying the rules, we need to create a table with the category id's of the products that our new customer has in his/her cart. As an example, we imagine that our customer has already introduced these items in the cart:

- Diet food.
- Milk, cheese and eggs.
- Fresh fruits and vegetables.
- Backed goods.
- Coffee, tea, cacao or tobacco.
- Frozen food and ice-creams.

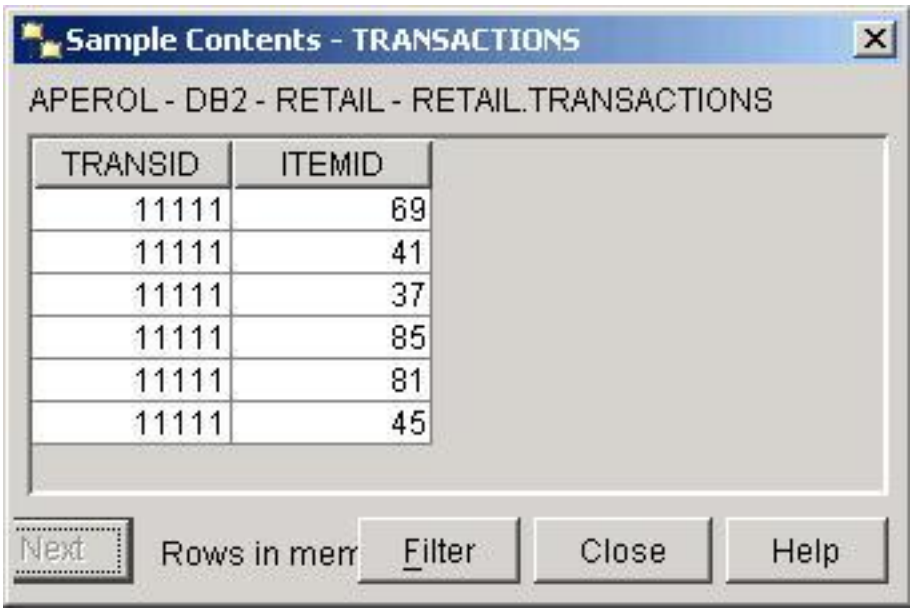
To create the table, execute:

```
[tutorial path]:\> db2 -tvf transactions.db2
```

You can click Troubleshooting on page 62 to check for errors that might occur when you use the above SQL command.

To view the content of this script, click *here*.

Here is the transactions table:



RETAIL.Transactions table: contains the categories id's of the products that are in the cart of our customer (his/her transaction id is 11111).

Deployment phase: how to apply the rules to the selected transactions

To apply the rules to the transactions, execute:

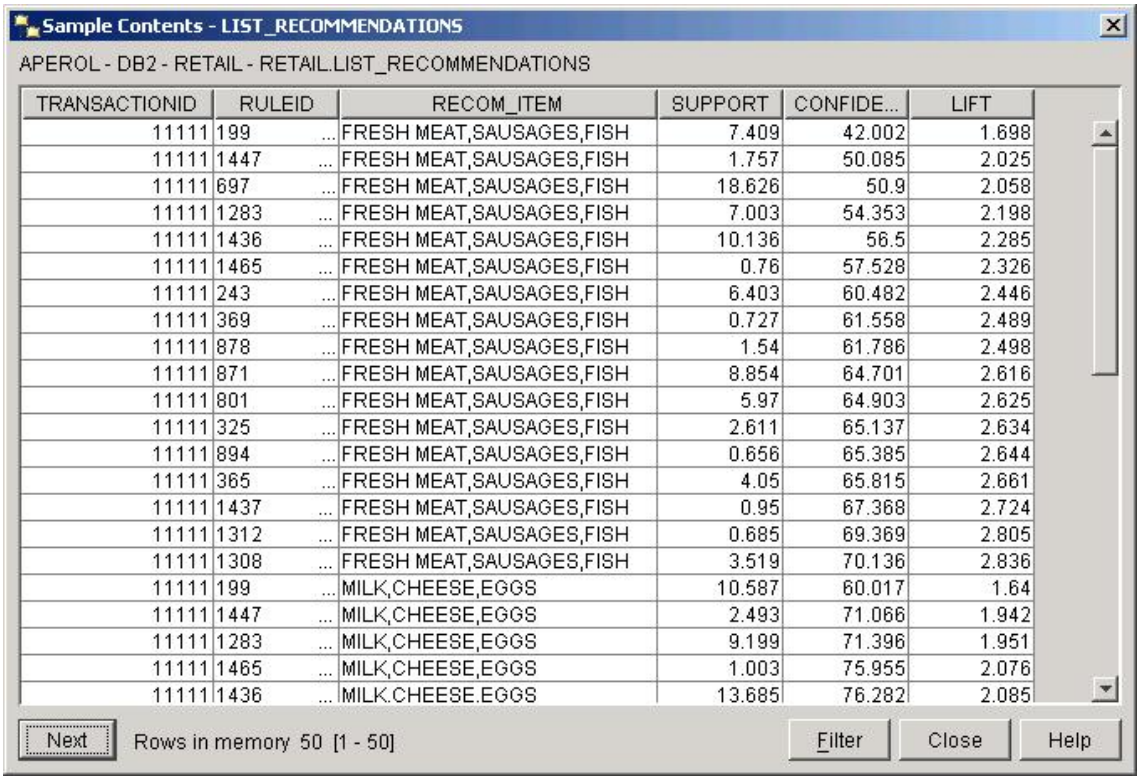
```
[tutorial path]:\>  db2 -tvf list_recommendations.db2
```

See Troubleshooting on page 62 to check for errors that might occur when you use the

above SQL command.

To view the content of this script, click *here*.

After the execution, we obtain a recommendation list table like this:


Sample Contents - LIST_RECOMMENDATIONS

APEROL - DB2 - RETAIL - RETAIL.LIST_RECOMMENDATIONS

TRANSACTIONID	RULEID	RECOM_ITEM	SUPPORT	CONFIDE...	LIFT
11111	199 ...	FRESH MEAT,SAUSAGES,FISH	7.409	42.002	1.698
11111	1447 ...	FRESH MEAT,SAUSAGES,FISH	1.757	50.085	2.025
11111	697 ...	FRESH MEAT,SAUSAGES,FISH	18.626	50.9	2.058
11111	1283 ...	FRESH MEAT,SAUSAGES,FISH	7.003	54.353	2.198
11111	1436 ...	FRESH MEAT,SAUSAGES,FISH	10.136	56.5	2.285
11111	1465 ...	FRESH MEAT,SAUSAGES,FISH	0.76	57.528	2.326
11111	243 ...	FRESH MEAT,SAUSAGES,FISH	6.403	60.482	2.446
11111	369 ...	FRESH MEAT,SAUSAGES,FISH	0.727	61.558	2.489
11111	878 ...	FRESH MEAT,SAUSAGES,FISH	1.54	61.786	2.498
11111	871 ...	FRESH MEAT,SAUSAGES,FISH	8.854	64.701	2.616
11111	801 ...	FRESH MEAT,SAUSAGES,FISH	5.97	64.903	2.625
11111	325 ...	FRESH MEAT,SAUSAGES,FISH	2.611	65.137	2.634
11111	894 ...	FRESH MEAT,SAUSAGES,FISH	0.656	65.385	2.644
11111	365 ...	FRESH MEAT,SAUSAGES,FISH	4.05	65.815	2.661
11111	1437 ...	FRESH MEAT,SAUSAGES,FISH	0.95	67.368	2.724
11111	1312 ...	FRESH MEAT,SAUSAGES,FISH	0.685	69.369	2.805
11111	1308 ...	FRESH MEAT,SAUSAGES,FISH	3.519	70.136	2.836
11111	199 ...	MILK,CHEESE,EGGS	10.587	60.017	1.64
11111	1447 ...	MILK,CHEESE,EGGS	2.493	71.066	1.942
11111	1283 ...	MILK,CHEESE,EGGS	9.199	71.396	1.951
11111	1465 ...	MILK,CHEESE,EGGS	1.003	75.955	2.076
11111	1436 ...	MILK,CHEESE,EGGS	13.685	76.282	2.085

Next Rows in memory 50 [1 - 50] Filter Close Help

List_Rules table: contains the rules generated by the RETAIL.ASSOC_MODEL2

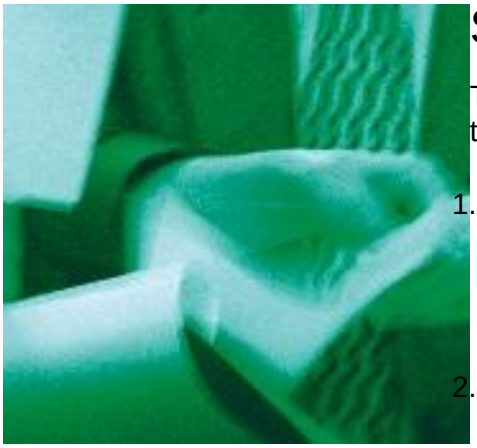
Here we see some of the products that would be recommended to our customer:

- Fresh meat, sausages and fish.
- Processed food (except soups).
- Soups, spices and spread products.
- Meat, sausages and fish tinned.
- Fruit, vegetables tinned.
- Beer and alcohol free drinks.
- ...

You probably have seen, that the recommended products appear more than one for the same transaction. This happens because we have come to this conclusion

(recommended product) by applying different rules. Actually, you can see the id of the rule that has been applied in the RULEID column.

On the other hand, to avoid harassing or stressing our new customer with so much products, we could only offer the top three that have a higher lift or support, for example. This could be an example of marketing strategy.



Summary

To create an associations model, you followed these steps:

1. In the see Data preparation phase on page 20 , you created the view RETAIL.ASSOC_VIEW. This table contains for each transaction, the category id of the item that was purchased in that transaction.
2. In the Modeling phase: building an associations model on page 24 , you built a model making a name mapping for the categories id, so that instead of obtaining rules with the category id, there was a match from the id to its description.
3. In the Evaluation phase: using IM Visualization on page 26 , you learnt how to open the IM Visualization to evaluate the model. First, you saw the rules that the model had generated and the statistics of the model (total number of rules, etc.). Afterwards, you learnt how to order the rules according to three important parameters: Support, Confidence and Lift.
4. Finally, in the Deployment phase: how to extract the rules into a table on page 33 , you saw how to generate product recommendations to the customers. First, we built a new model without name mapping. Then, we extracted the rules into a table. To apply the rules to some transactions, we invented a list of items that our customer had on his/her cart and created a new table (RETAIL.TRANSACTIONS) with that information. Then, the last step was to apply the rules to the transactions and obtain the results.

Exercises

We strongly recommend you to try to solve this exercises to strengthen the concepts learned in this section.

Answer the following questions:

1. What can we obtain with the ASSOCIATIONS mining technique? Answers to Question 1 on page 39

2. What Data Preparation do we need in the ASSOCIATIONS? Answers to Question 2 on page 40

3. How can we obtain less and better rules? Answers to Question 3 on page 41

You can skip the exercises and go to the next section Introducing the clustering mining function on page 42 .

Answers to Question 1

What can we obtain with the ASSOCIATIONS mining technique?

With the Associations mining function we can obtain a model with a set of rules that explain the linking purchasing behaviour of the customers. This means, that we can have rules such as: A + B => C, that is to say, if a customer buys the item A and the item B, it is very likely that he buys the item C as well. The likelihood is determined by the Confidence and Support values of the rule.

With the rules, we can generate product recommendations for future clients, taking into account what they have already purchased, and the linking or associations rules that we have from the model.

You can skip the exercises and go to the next section Introducing the clustering mining function on page 42 .

Back

Answers to Question 2

What Data Preparation do we need in the ASSOCIATIONS?

To apply the associations mining technique in our scenario, we only need to gather in one view or table the information needed. In our case, since the Retail company wants to get rules that link products that were bought in the same transaction, we only need a table or a view with two columns. The first column is the transaction id and the second the item purchased.

You can skip the exercises and go to the next section Introducing the clustering mining function on page 42 .

Back

Answers to Question 3

How can we obtain less and better rules?

First, we have to define what are better or good rules. As a general guideline, the best rules are those with a higher Lift, Confidence and Support. When you build a model and write no constraints for the minimum values that theses parameters should have, you probably obtain a big amount or rules. What you can do is order them by this parameter values, clicking on the columns in the IM Visualizer, as we saw it in the evaluation phase.

You can also rebuild the model setting higher constraints for the values of the Lift, Confidence and Support, so that all the rules you obtain fulfill the minimum requirements you set. This way, you will obtain less rules. In fact, this is a way of dismissing those rules that are not very interesting.

You can skip the exercises and go to the next section Introducing the clustering mining function on page 42 .

Back

Section 5. Creating a clustering model

Introducing the clustering mining function

The Clustering belongs to the segmentation Data Mining techniques, which are mainly used for customer segmentation into groups or clusters. Indeed, there many techniques for clustering, but among all of them, in this scenario we will use the IBM DB2 Intelligent Miner Demographic clustering mining techinques.

Demographic Clustering

the number of clusters (or segments or groups) are automatically decided based on the measure representing how similar the records within the individual clusters should be. However, you can indicate the maximum number of clusters that you want to obtain. Generally, it is used when most of the variables are categorical.

It also requires you to specify how similar the customers within clusters should be - the **similarity threshold** parameter. The threshold ranges between 0 and 1, where value of 1 means that all the customers in a cluster must be identical, and value of 0 means that customers can be completely different. So if we set the threshold to 1, we will probably will obtain as much clusters as customers because we would not find two customers that were 100% identical to be grouped in the same cluster. On the other hand, if the threshold is set to 0, we would only obtain 1 cluster with all the clients.

As a general rule, if two customers are more similar than the threshold value, then they are candidates for being put into the same cluster. We say candidates because it could also happen that there was another group of customers (cluster) where the match was better.

The demographic clustering mining technique finds optimum combinations of customers that maximize their similarity within each cluster, while at the same time maximizing the dissimilarity between different clusters.

Data preparation phase: Introduction

Here, as with the Associations technique, we will collect the data (tables, views) which will help us to obtain customer segmentation. In the Data preparation phase we can check and delete invalid values (outliers, missing values, etc.) to clean and format the data. Furthermore, and if necessary, we can even create new attributes, such as: averages, derived attributes, etc.

For the Clustering in our scenario, we will use one of the tables that we used when applying Associations as raw material. This table was the *RETAIL.POS_DATA* and it contains the data from only one store in one month. Among all, we only take those that have a CUSTOMERID different from 0, which means that they have a customer card. Moreover, we will discard the customers that have purchased less than 20 items, because we need a minimum of shopping activity for our analysis.

However, this will not be enough. To characterize our customers, we want to know a little bit more about their purchasing behaviour. Therefore, we need to know WHEN do they go shopping and WHAT do they shop (quantity/product, revenue spent/product). Because there are lots of different item categories defined in the Articles table, we should group them into bigger groups or families, such as:

- FRESH_PRODUCTS, which includes fresh meat, fish, vegetables and fruits.
- MILK_CHEESE_EGGS
- FROZEN_FOOD, which includes ice creams, meat, fish, pizzas, etc.
- PROC_TINNED_FOOD, which includes processed food like soups and all kinds of tinned food (meat, vegetables and fruits).
- BACKED_CONFECTIONERY, which includes bread and cakes.
- DIET_FOOD
- ALCH_DRINKS, all kind of alcohol drinks.
- NON_ALCH_DRINKS_TOBACCO, like coffee, tea, etc. and tobacco.
- DET_HYG_COSMETICS, which contains cleaning, hygienic and cosmetic products.
- HOME_TEXTILES_DECO, all kind of textiles for the house (towels, curtains, etc.) and decoration products like porcelain, etc.
- CLOTHES_UNDERWEAR
- BABY_PRODUCTS
- COMPL_SHOES_LEATHER, includes complements like ties, belts, etc, shoes and leather clothes.
- SPORT, all sport products (clothes and devices).
- CAMPING_GARDEN_FURN, includes furniture for the garden and camping material.
- GAMES_JOKES_PARTY.
- STATIONERY, including paper, books and school articles, among other.
- ELECTRONICS, entertainment, foto, video, computer, etc.
- BUILDING, as building material, wallpapers, etc.
- CAR_BIKE_ACC, car and bike accessories.
- PETS_GARDEN, pet food and complements and garden products like plants, earth, etc.
- OTHER

For each family listed above we will create two derived attributes:

- one for the QUANTITY of items purchased that belong to that family. (it will be named Q_nameattribute)
- one for the total REVENUE spent in the items of that family. (it will be named R_nameattribute)

This quantities will be given per customer and in percentage values, that is, R_itemfamily (relative) = R_itemfamily / R_totalrevenuespent * 100. The same for the quantities: Q_itemfamily (relative) = Q_itemfamily / Q_totalquantity * 100. Finally, we have also created derived attributes that indicate the revenue spent and the times that each customer went shopping in each weekday (including Saturday).

Data preparation phase: How to prepare the data

The input data (Glossary on page 63) for the mining function will be a view or table that has all those attributes which we want to take into account to segment the customers. To build a table with all the attributes takes some time in this case. Execute the following command:

```
[tutorial path]:\> db2 -tvf clustering_preparation.db2
```

To view content of this script, click *here*.

Here are the summarized steps that the script executes to prepare the data:

1. Creates a table (RETAIL.ARTICLE_FAMILIES) with the description of the new families we specified.
2. Creates a table (RETAIL.FAMILY_MATCHING) that matches each category id to a familyid.
3. Creates a view (RETAIL.CUSTOMER_DATA) with the customerid, the transid, the item family description, the date, the time, the quantity and the price.
4. Creates a view (RETAIL.CUSTOMER_DATA2) with the total amount of items and revenue spent per customer and family product. this table has 4 attributes: CUSTOMERID, ITEM (family description), QUANTITY and REVENUE.
5. Creates a table (RETAIL.CUSTOMERS) with all the information in the prior view, we create a table with the CUSTOMERID and 2 columns per family item. The first will refer to the revenue spent in that family of products and the second to the quantity of products that were purchased for that family. Furthermore, there will be some fields for the quantity and revenues for each weekday and some attributes that will contain the the total amount of revenues, times that each customer went shopping (transactions) and amount of items he/she purchased (quantity).

6. To fill the table we have to make a set of full outer joins to count, for each family and customer, the revenues spent and quantity purchased.
7. View (RETAIL.DAYS)to translate the date from the RETAIL.POS_DATA table into a day of the week.
8. View (RETAIL.FREQUENCY) to count how many times and how much each customer spent.
9. View (RETAIL.FREQUENCY_CUSTOMERS2) that counts per weekday (from Monday to Saturday) the quantity and revenue spent
10. View (RETAIL.FREQUENCY_CUSTOMERS4) to make the make the values (times and quantity) relative to the total times and quantity.
11. Inserts all the data we have into the RETAIL.CUSTOMERS table.
12. Updates the values of each family to make them relative to the total revenue spent and total times that each customer went shopping.

You can click Troubleshooting on page 62 to check for errors that might occur when you use the above SQL command.

After the execution, you will obtain the following table:

Sample Contents - CUSTOMERS

APEROL - DB2 - RETAIL - RETAIL.CUSTOMERS

CUSTOMERID	R_FRESH...	Q_FRESH...	R_MILK_C...	Q_MILK_C...	R_FROZE...	Q_FROZE...	R_PROC_...	G
7,968	9.036	20	0	0	0	0	0	
8,258	0	0	0	0	0	0	0	
8,502	5.615	11.111	0	0	1.218	2.222	8.422	
8,509	0.784	2.703	0	0	0.353	2.703	0	
8,526	0	0	0	0	0	0	0.423	
8,530	3.172	4.516	0	0	3.417	7.742	7.66	
8,533	7.031	12.5	0	0	0	0	1.188	
79,499	0	0	0	0	0	0	0	
80,226	2.225	2.857	0	0	0	0	1.718	
80,498	0	0	0	0	0.864	3.846	0	
82,478	1.661	9.756	0	0	1.036	2.439	0.404	
82,494	3.294	5.556	0	0	0	0	2.817	
82,551	0	0	0	0	0	0	0	
85,088	0.795	2.326	0	0	0	0	7.24	
85,264	0.356	1.818	0	0	1.011	3.636	0.113	
962,660	0	0	0	0	4.291	17.857	1.708	
9,616,976	9.272	8	0	0	2.537	4	15.743	

Next Rows in memory 50 [1 - 50] Filter Close Help

Customers table: contains all the information related to the customers, as Customer_id, quantity and revenue spent

in each item family and the times that each customer has gone shopping on Monday,

Tuesday, Wednesday, Thursday,

Friday and Saturday.

Modeling phase: using Easy Mining procedures

In this section will present how easy it is to build CLUSTERING models with the Easy Mining Procedures of the IBM DB2 Intelligent Miner.

Here again, there are many procedures to build clustering models, which can be found in the the *ClusProcsDecl.sql* file. We will focus on the following procedure call:

```
IDMMX.BuildClusterModel(<modelName>, <tableName>, <optionalParameters>)
```

Where:

- <*modelName*> is the name that we will give to the model built.
- <*tableName*> is the table or view that contains the Data Model (CUSTOMERS for the clustering in the Retail case).
- <*optionalParameters*>, among other, can be:

1. *' '*. Empty, when no settings or optional parameters are required.
2. *DM_setMaxNumClus(<max_num_clusters>)*: to specify the maximum number of clusters that we want to obtain. Note, that with Demographic Clustering, the clusters will be built automatically, we cannot specify how many do we want to obtain. However, we can specify the maximum number we want to obtain.
3. *'DM_remDataSpecFld(<item>)'*: to remove some attributes from the 'tableName' table, that we DO NOT want to take into account for the clustering. For example, in the CUSTOMERS table, if we did not want to apply the clustering mining function to segment our customers depending on the revenue they spent per week day, we would do: DM_remDataSpecFld of (MO_REVENUE, TU_REVENUE, ... SA_REVENUE).

The output of this procedure call will be a CLUSTERING model that will be stored in an output table, the IDMMX.CLUSTERMODELS table.

Modeling phase: building a clustering model

In the retail scenario, we can call:

```
[tutorial path]:\> db2 -tvf buildclusmodel.db2
```

Troubleshooting on page 62

The content of the script is:

```
call IDMMX.BuildClusterModel('RETAIL.CLUS_MODEL', 'Retail.CUSTOMERS', ' ');
```

This call will generate a clustering model for the CUSTOMERS table. We do not give any extra optional parameters to get a first approach into the clustering models.

Now that the model is stored in the (IDMMX.CLUSTERMODELS table), we will use the IBM DB2 Visualization to open it and evaluate it. This will be the next phase: the evaluation.

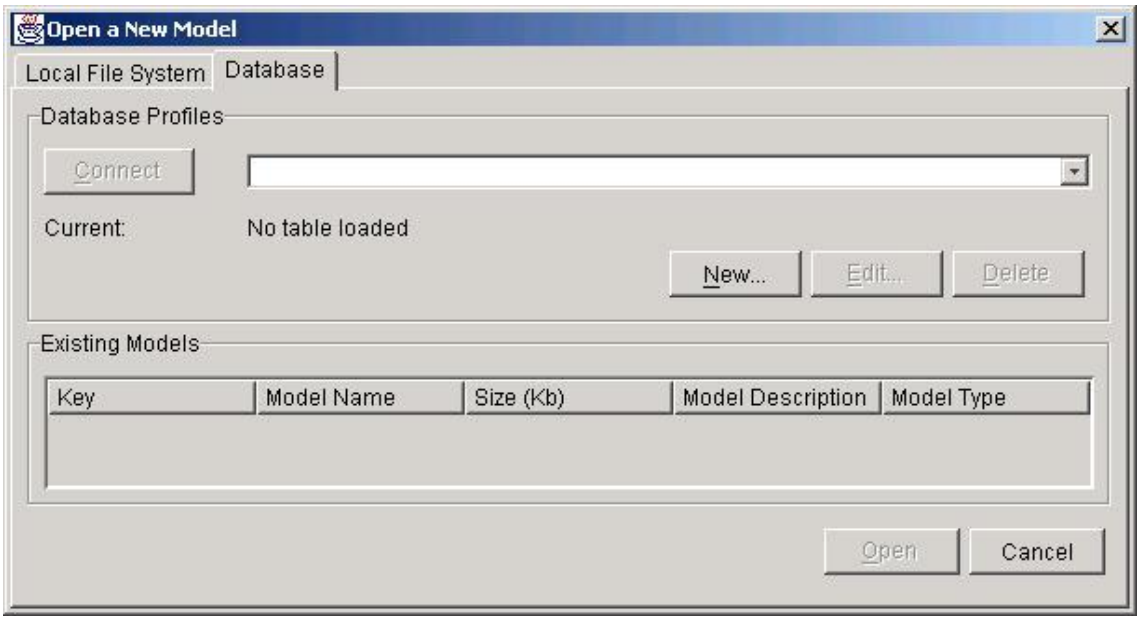
Evaluation phase: using IM Visualization to asses the model

When the model is created, you can use IBM DB2 Intelligent Miner Visualization to look at the results and to evaluate whether the model is good for your business problem.

Connecting to a database

Note: We assume that at this point you have already been through the Evaluation of the associations model and therefore, copied the *db2java.zip* and *db2jcc.jar* files.

When you start the IM Visualizer, click the **Database** label of the window displayed:



The Open a New Model window

On the Open a New Model window, you can select one of the models in the *Existing models* list. If you use the IM Visualizer for the first time, the Existing Models list is empty. Before you can view the exiasting models, you must create database profiles.

To create a new database profile, click **New...**.

Creating a new database profile

On the Add New Database Profile window, type the following parameters in the appropriate entry fields and click **Retrieve Database Information**:

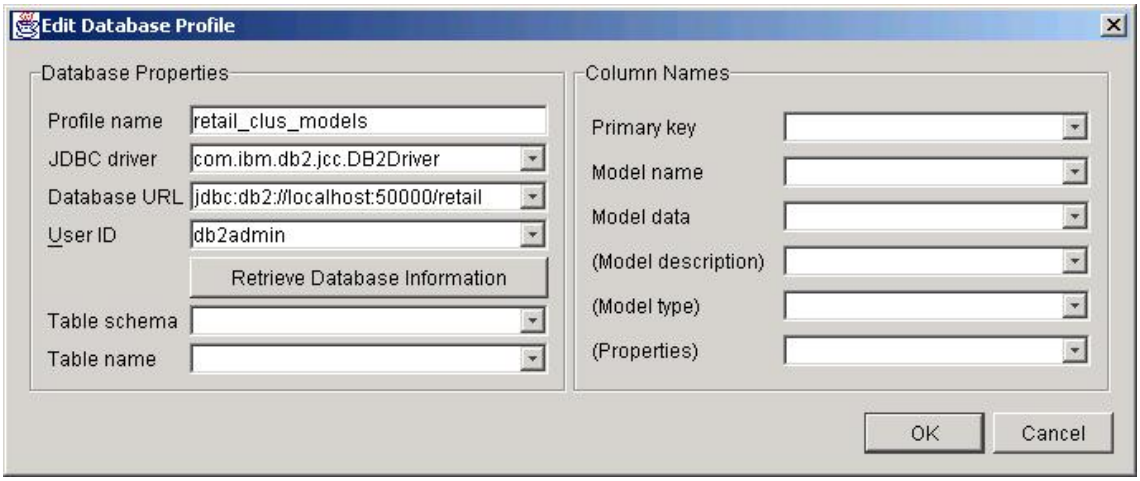
Note: The name of the JDBC driver is automatically displayed, but make sure that corresponds to the one specified below. If you cannot create the Database Profile, it might be that your JDBC driver and/or you URL are different. In this case, contact your system administrator.

Profile name retail_clus_models

JDBC driver com.ibm.db2.jcc.DB2Driver

Database URL jdbc:db2://localhost:50000/retail (localhost if DB2 was localy installed in your machine. Otherwise, contact your system administrator).

UserID db2admin (for example)



The Add New Database Profile window

Type the following parameters in the appropriate entry fields and click **OK** to return to the Open a New Model window: Table schema IDMMX

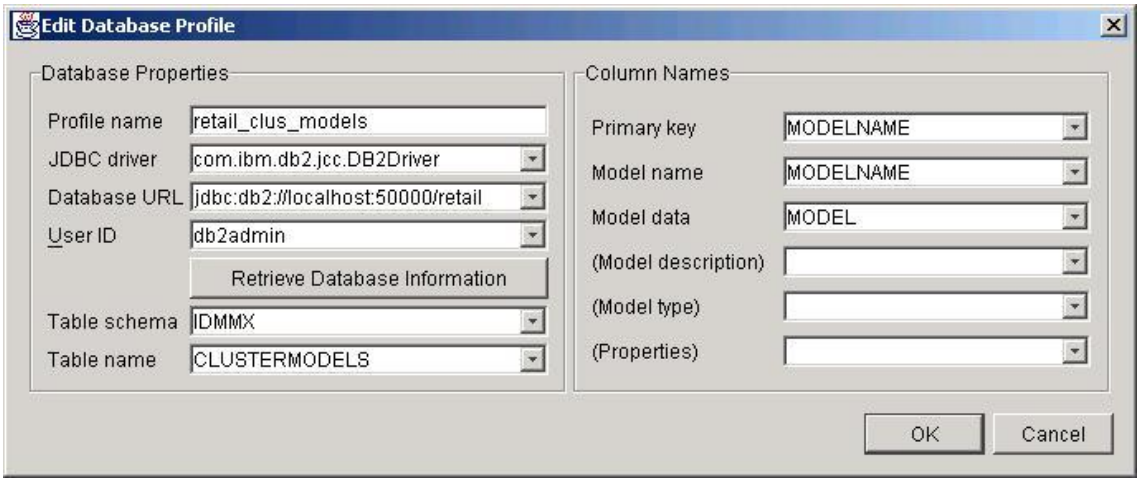
Table name CLUSTERMODELS

Primary key MODELNAME

Model name MODELNAME

Model data MODEL

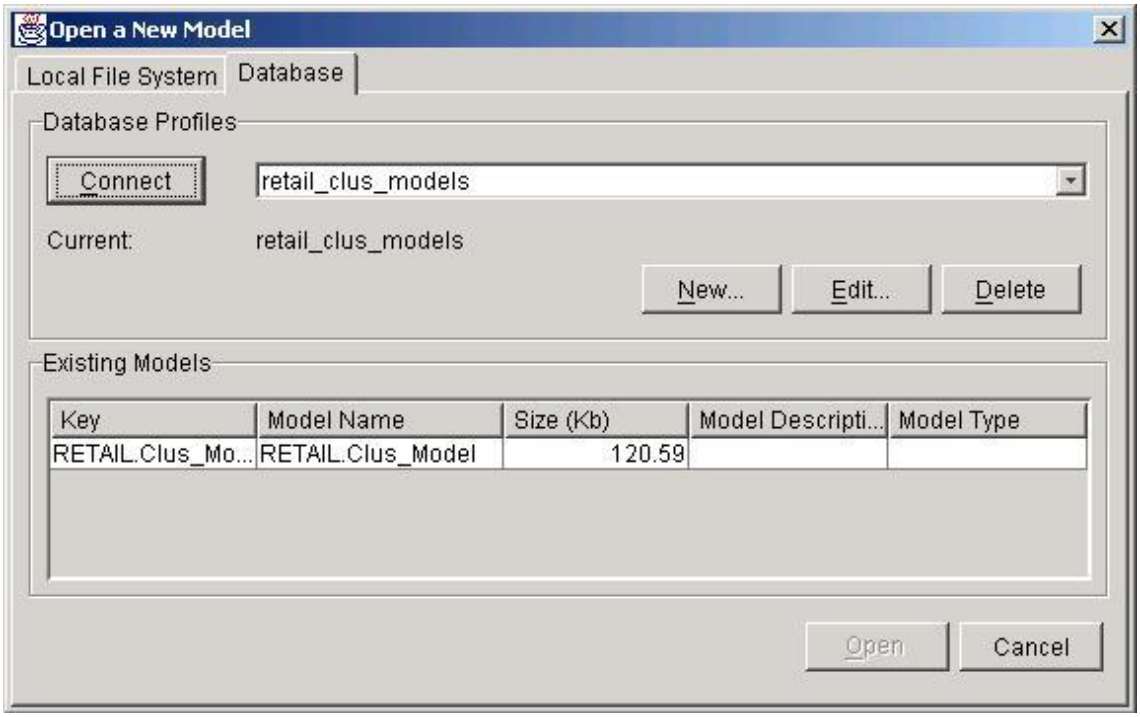
The following graphic shows the Add New Database Profile window with the complete specification:



The complete specification of the Add New Database Profile window

Opening a model

On the Add New Database Profile window, click **Connect** to connect to the database that you have specified on the previous window. When your operating system is connected to the database, a list of models that are available in this database is displayed.

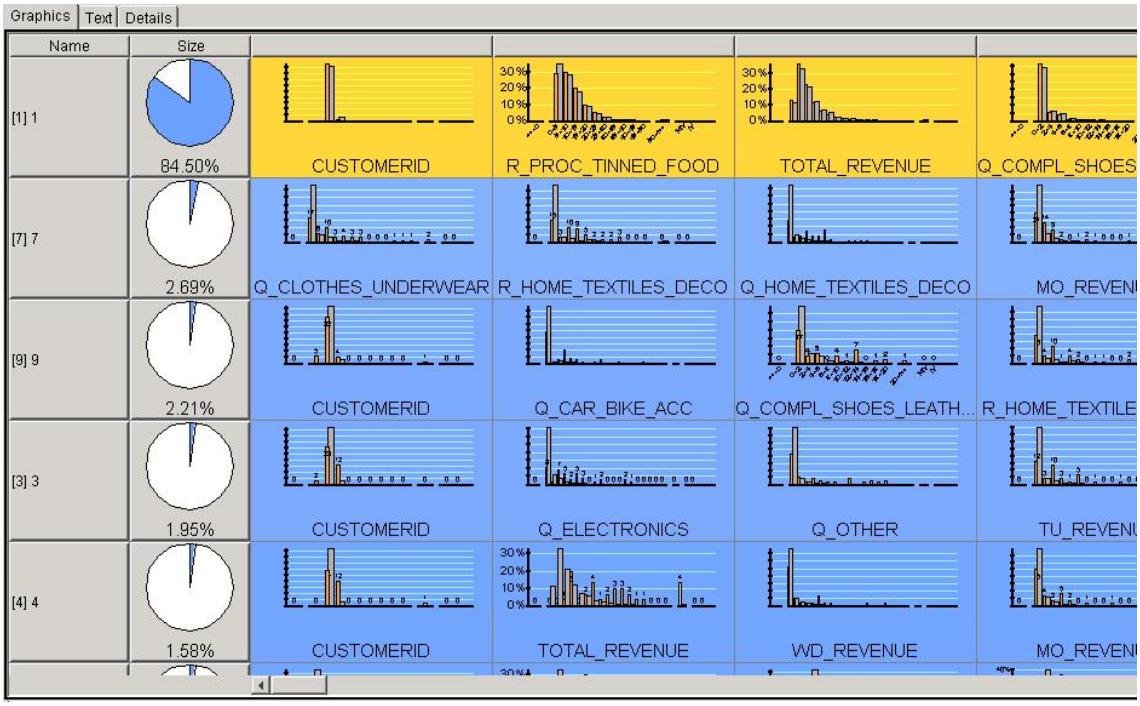


The Open a New Model window displaying a list of existing models

To open the RETAIL.CLUS_MODEL, select this model from the list of existing models and click **Open**.

Evaluation phase: interpreting the results

When you are first confronted with the cluster results the first question that you are going to ask is what does it al mean. In this section we describe how to understand and read the visualized segmentation output and how to interpret the results.



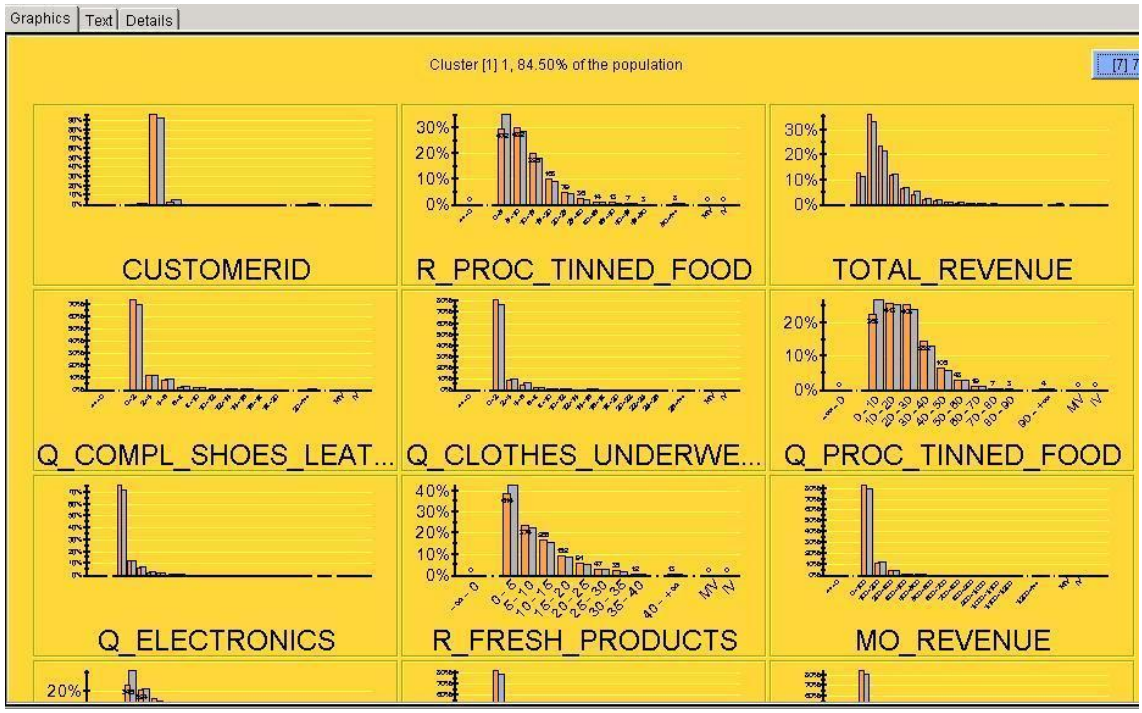
The results of the model showed with the IBM DB2 IM Visualization: the different segments that the model calculated

and its characteristics are presented.

We can observe the different clusters or segments that the model has built. The first column indicates the name or code given to the cluster, while the second represents the percentage of the population that belongs to that cluster. From then on, you can see the distribution that the customers in each segment follow for each attribute.

To see more clear the characteristics of each cluster, double click on it. To go back to the previous view of the model, click the back button (left arrow) in the toolbar.

These are the characteristics of the first and bigger cluster (84.50%):



The characteristics of the customers classified into this cluster.

In each diagram, the gray bars represent the distribution of all the customers in our Database, whereas the orange ones the distribution of the customers that belong to the first cluster (the one we are analyzing). If we compare both distributions, we can observe that the customers that belong to this cluster, spent less than the average customers. You can see this in the distribution of the TOTA_REVENUE table. Actually, you can double click the diagram to see it better. Afterwards, you can click the left arrow in the menu to see the previous view of the model.

However, **this model is not good enough**. Here is the justification:

1. CUSTOMERID was used to segment the customers: as every customer has a different CUSTOMERID, it provides no information if we use it to segment.
2. Some attributes like Q_BABY_PRODUCTS and R_BABY_PRODUCTS or Q_PETS_GARDEN and R_PETS_GARDEN, among other, do not have different distributions from the general distribution, no matter the cluster we analyze. This means that these attributes do not provide important information to segment the customers. It is clear that if all the customers follow the same distribution for an attribute, there is no use in using it to classify the customers in different groups because they will be matched in the same group.

For all these reasons the model cannot be considered as a good one. In our next step,

we will go back in the Data Mining Process to the modeling phase. We will build a new model with some constraints about the parameters or attributes, so that we can ignore those that bring no information.

Modeling phase: building a BETTER clustering model

Now, we are interested in building a new model that does not take into account:

In our scenario this attributes were:

- R_BABY_PRODUCTS, Q_BABY_PRODUCTS, R_PETS_GARDEN, Q_PETS_GARDEN, R_MILK_CHEESE_EGGS, Q_MILK_CHEESE_EGGS, R_OTHER, Q_OTHER, R_CAR_BIKE_ACC, Q_CAR_BIKE_ACC, R_BUILDING, Q_BUILDING, R_CAMPING_GARDEN_FURN and Q_CAMPING_GARDEN_FURN, R_SPORT, Q_SPORT, R_GAMES_JOKES_PARTY, Q_GAMES_JOKES_PARTY, R_DIET_FOOD, Q_DIET_FOOD.

To build a new clustering model, execute:

```
[tutorial path]:\> db2 -tvf buildclusmodel2.db2
```

You can click Troubleshooting on page 62 to check for errors that might occur when you use the above SQL command.

The content of the script is:

```
CONNECT TO RETAIL;

call IDMMX.BuildClusModel( 'RETAIL.CLUS_MODEL2', 'RETAIL.CUSTOMERS',
'DM_setMaxNumClus(9), DM_setDClusPar(''SimThr'',0.6),
DM_remDataSpecFld(''CUSTOMERID''),
DM_remDataSpecFld(''R_BABY_PRODUCTS''),
DM_remDataSpecFld(''Q_BABY_PRODUCTS''),
DM_remDataSpecFld(''R_PETS_GARDEN''),
DM_remDataSpecFld(''Q_PETS_GARDEN''),
DM_remDataSpecFld(''R_MILK_CHEESE_EGGS''),
DM_remDataSpecFld(''Q_MILK_CHEESE_EGGS''),
DM_remDataSpecFld(''R_OTHER''),
DM_remDataSpecFld(''Q_OTHER''),
DM_remDataSpecFld(''R_CAR_BIKE_ACC''),
DM_remDataSpecFld(''Q_CAR_BIKE_ACC''),
DM_remDataSpecFld(''Q_BUILDING''),
```

```
DM_remDataSpecFld(''Q_HOME_TEXTILES_DECO''),
DM_remDataSpecFld(''Q_COMPL_SHOES_LEATHER''),
DM_remDataSpecFld(''Q_CAMPING_GARDEN_FURN''),
DM_remDataSpecFld(''Q_SPORT''),
DM_remDataSpecFld(''R_SPORT''),
DM_remDataSpecFld(''Q_DET_HYG_COSMETICS''),
DM_remDataSpecFld(''Q_FRESH_PRODUCTS''),
DM_remDataSpecFld(''Q_FROZEN_FOOD''),
DM_remDataSpecFld(''Q_ELECTRONICS''),
DM_remDataSpecFld(''Q_STATIONERY''),
DM_remDataSpecFld(''Q_CLOTHES_UNDERWEAR''),
DM_remDataSpecFld(''Q_PROC_TINNED_FOOD''),
DM_remDataSpecFld(''Q_DIET_FOOD''),
DM_remDataSpecFld(''Q_BACKED_CONFECTIONERY''),
DM_remDataSpecFld(''Q_GAMES_JOKES_PARTY''),
DM_remDataSpecFld(''Q_ALCH_DRINKS''),
DM_remDataSpecFld(''Q_NON_ALCH_DRINKS_TOBACCO''),
DM_remDataSpecFld(''R_CAMPING_GARDEN_FURN''),
DM_remDataSpecFld(''R_BUILDING''),
DM_remDataSpecFld(''R_ELECTRONICS''),
DM_remDataSpecFld(''R_GAMES_JOKES_PARTY''),
DM_remDataSpecFld(''R_DIET_FOOD''),
DM_remDataSpecFld(''MO_TIMES''),DM_remDataSpecFld(''TU_TIMES''),
DM_remDataSpecFld(''WD_TIMES''),DM_remDataSpecFld(''TH_TIMES''),
DM_remDataSpecFld(''FR_TIMES''),DM_remDataSpecFld(''SA_TIMES'')
');
CONNECT RESET;
```

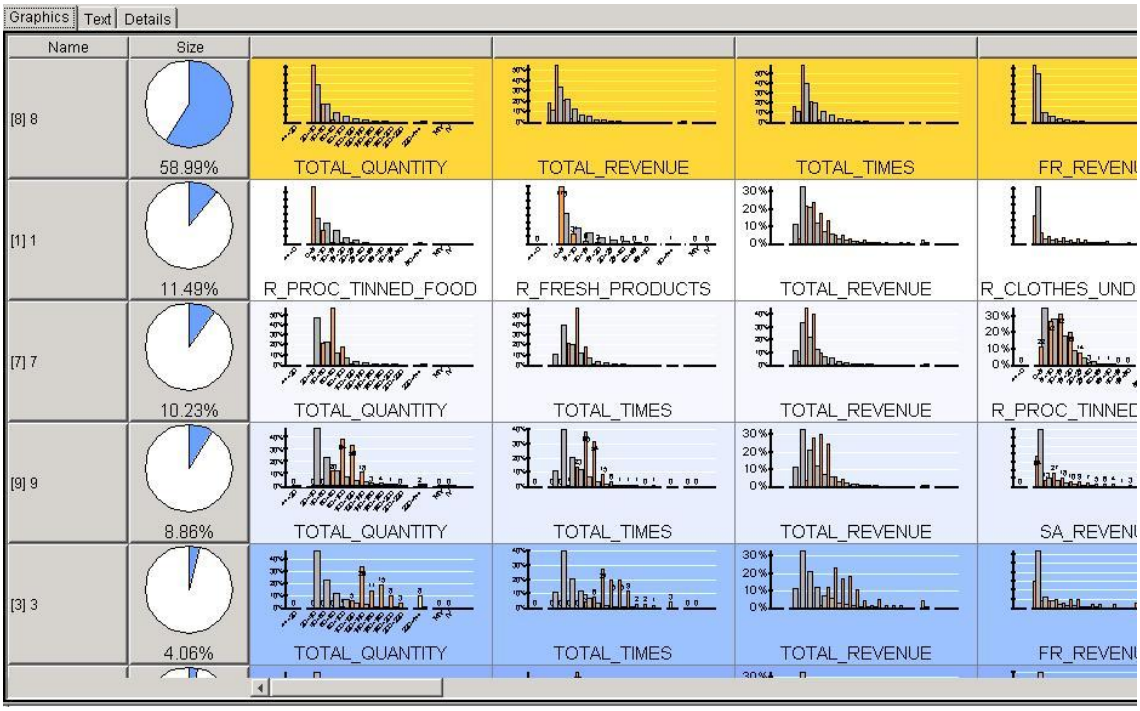
This call will generate a clustering model for the CUSTOMERS table without the specified attributes. Furthermore, we have set the maximum number of clusters to 9 and the similarity threshold to 0.6. This way we guarantee that our model will not have more than 9 clusters and that the similarity between any two customers that belong to the same cluster is 0.6 so that we can obtain more differenced clusters.

Again, the model was stored in the Database (IDMMX.CLUSTERMODELS table). Let's open it and evaluate its results with the IM Visualization.

Evaluation phase: interpreting the new results

Open the new model, RETAIL.CLUS_MODEL2, with the IM Visualization, as we did with our first model. In this case, you only have to open the IM Visualization and connect to the 'retail_clus_models' database profile and then click on the 'RETAIL.CLUS_MODEL2' to open it.

You will see these results:



The results of the new model.

We can observe that now the those attributes that we removed are no longer presented. In our next step, we will analyze each attribute, to define the characteristics of each cluster. This way, we will understand better the customers that the Retail Company has.

Evaluation phase: analyzing the clusters

To understand better what kind of customers the Retail Company has, we will analyze the distributions of the attributes in each cluster. Actually, although we have obtained 9 Clusters, we will only analyze the the 5 biggest, as an example. You can double click each cluster to see all the distributions of the variables and/or click the second label (Text), to get a description of the distribution of the attributes values for each cluster.

FIRST CLUSTER: (named 8, 58.99%)

If we focus on the first and most populated cluster, we see that the customers that belong to it are characterized by purchasing less than the general distribuition. However, as this is the bigger cluster, we can consider the customers in it as **GENERAL CUSTOMERS**.

SECOND CLUSTER: (named 1, 11.49%)

In the second, customers that belong to it are characterized by spending more money in their shoppings. Moreover, they come to buy not only on Fridays and Saturdays but also on Monday and Tuesday. Among their favourite products, they purchase more clothes and underwear, complements, shoes and leather, etc. On the other hand, they purchase less food products such as processed food, tinned food and fresh products, etc.

So we can say that the customers belonging to the second cluster are weekday customers (specially on Monday) and they mainly buy non-food products. For these reasons we could catalogue them as **WEEKDAYS NON-FOOD CUSTOMERS**.

THIRD CLUSTER: (named 7, 10.23%)

The third cluster is mainly integrated by customers that spend a lot of money in their shopping, which is mainly based on food products. The total quantity of products bought is quite high and despite they go buying any day of the week, they prefer Fridays.

Customers belonging to the third cluster are **LOYAL FOOD CUSTOMERS**.

FOURTH CLUSTER: (named 9, 8.86%)

In the fourth cluster, the customers purchase detergents, hygienic products and cosmetics but also food and non-food products. They go shopping specially on Friday and Saturdays.

The customers of this cluster can be named as **LOYAL WEEKEND CUSTOMERS**.

FIFTH CLUSTER: (named 3, 4.06%)

The fourth cluster is mainly integrated by customers that buy big amounts of products and not only food, but also detergents, clothes and underwear, etc. Regarding when do they go shopping, Friday and Thursdays seem to be their favourite days, although they also spent a lot on the rest of the days.

The customers of this cluster can be named as **LOYAL GENERAL CUSTOMERS**.

All in all, we see that there are 5 main categories of clients that have different purchasing behaviours. It would be wise for the retail company to exploit this information and generate specific campaigns for each customer, depending on to which segment does he/she belong. This will be done in the last step of the Data Mining Process, the deployment.

Deployment phase: exercise

As deployment, now you can make better product recommendations to the same customer that we used for the association. Our customer was adding new products to the cart (RETAIL.TRANSACTIONS table) and seeing possible recommendations in a display. Now instead of recommending him products, considering the purchase behaviour of all the clients, we can give better recommendations only taking into account the purchasing behaviour of customers that are similar to him.

Note: to do this exercise we will apply the clustering model using idmmx.applyClusModel. You will need the PTF (programme temporal fix) for this, which will be released in the third quarter of 2003.

We assume that we already know which is his cluster, for example the third one (LOYAL FOOD CUSTOMERS). Actually, the finding out of his cluster could be done through a customer card or some kind of identification mechanism. Yet, your task now is to make recommendations according to behaviour of other customers similar to him (in the same cluster). To do this, first you apply clustering and then associations to build specific rules for the customers in the fourth cluster. With this exercise you will remember better the techniques shown in this tutorial.

The steps to be followed will be given as a guide and you can find the solution in a script (*solution.db2*).

Steps to be done:

1. Apply the clustering model to the customers: this way you classify them into the different clusters. You have to use the IDMMX.applyClusModel function which belongs to the IM Scoring functions because it applies the model to new data. The function is defined below:

```
IDMMX.ApplyClusModel (<modelName>, <tableName>,
                      <outputView>, <clusterIdColumn>,
                      <qualityColumn>, <confidenceColumn>)
```

   Where: <modelName> is the name of the model we apply.

('RETAIL.CLUS_MODEL2', in our case) <tableName> is the name of the input table with the customers. ('RETAIL.CUSTOMERS', in our case) <outputView> is the name of the view where we will obtain the results.('Retail.CLUSTERS', for example) <clusterIdColumn> is the name we want to give to the column that will indicate to which cluster belongs each customer. ('Cluster', for example) <qualityColumn> is the name we want to give to the column that will indicate how well the applied record fits into the specified cluster.('Quality', for example). <confidenceColumn> is the name we want to give to the column that will indicate the confidence that the customer belongs to the cluster/segment to which they has been assigned. ('Confidence', for example)

2. Get from the output view the customers that belong to the cluster we want to analyze. We take the third cluster (named 7), for example. We can do this with a simple sql statement that creates a view as a subselect of the output view RETAIL.CLUSTERS.
3. Get the transactions for the customers that belong to the cluster.
4. Apply the associations model to the transactions of the customers.
5. Extract the output rules into a table.
6. Apply the rules to the items that the new customer has added to his cart.

NOTE: if you want to execute the solution to this exercise, do:

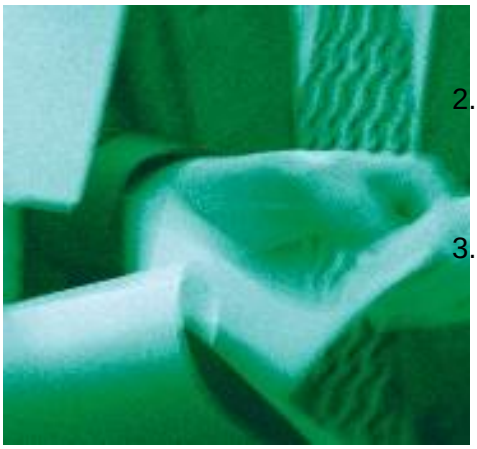
```
[tutorial path]:\> db2 -tvf solution.db2
```

The output view is called RETAIL.List_Recommendations_Cluster7.

Summary

To create a clustering model, you followed these steps:

1. In the Data preparation phase: How to prepare the data on page 44 , you created a new table CUSTOMERS that gathered all the information for each customer, as: relative revenue and relative quantity spent in each family of products. Moreover, there was also some columns that indicated how many times did he/she go



shopping per weekday and how much revenue he/she spent.

2. In the Modeling phase: building a clustering model on page 47 , you learnt how to build a clustering model with the IBM Easy Mining procedures.
3. In the Evaluation phase: interpreting the results on page 50 , you saw how to open and evaluate the model with the IBM DB2 IM Visualization. The first model we created was not good, as it was taking into account attributes such as the CUSTOMERID to segment the clients.
4. You went back in the Data Mining process to build a better model (Modeling phase: building a BETTER clustering model on page 53), removing those attributes that brought no extra or important information, or that could simply disturb the results. You learnt how to do it, specifying it as optionalParameter of the Easy Mining procedure.
5. In the evaluation of the new model (Evaluation phase: interpreting the new results on page 54), you learnt how to analyze each cluster and understand what kind of customers does the Retail Company have.
6. Last but not least, as a deployment (Deployment phase: exercise on page 57) we encouraged you to make an exercise. The challenge was to obtain product recommendations for a specific cluster, that is, for the customers that belonged to one cluster. The solution of the exercise is attached in this tutorial as a sql script: solution.db2. To see the final results you can execute it or simply open it, to see how we did it.

Section 6. Application integration

Integrate data mining into your corporate applications

It would be interesting for the retail company to observe the rules over several months. Changes could be seen in the behaviour of the consumers. This could be one of the possible integrations of the Easy Mining prodcedures into corporate applications. In this case, a problem comes up when the amount of data to be mined is so huge that a single mining task might take hours to be finished. Therefore, it is very interesting to see how you can build a web application that handles multiple mining tasks in the background. Hence, we recommend you to read the *Building a data mining solution using IBM Intelligent Miner Modeling and Websphere* tutorial to learn how to do it. The web application built also provides a GUI that allows you to start, cancel and delete mining tasks, as well as view the results.

On the other hand, following with the product recommendation, the managers of the retail company would be interested in using real-time scoring to personalize their customer contact. In this sense, we recommend you to have a look at the *Building a data mining solution using IBM Intelligent Miner Scoring and Websphere* tutorial, which shows you how to provide advanced personalized self-service offerings on your website.

Section 7. Tutorial Summary

Mining your Business in Retail with IBM DB2 Intelligent Miner

We have learned how to generate explanations for our Retail Company scenario using two mining techniques: ASSOCIATIONS and CLUSTERING.

For both techniques we have had to create a new views and tables that gathered all the necessary information. We saw it in the Data Preparation Step. With clustering furthermore, we had to derive a lot of new attributes, that defined product families and the revenue that was spent and the quantity of products purchased for each family.

To build the models, we used the IBM Easy Mining procedures for associations and clustering. Then we opened and evaluated them with the IM Visualization. Sometimes, it was necessary to rebuild the model, when we considered that it was not good enough. Evaluating the models we could answer the business questions regarding what is the purchasing behaviour of the Retail Company's customers (the rules indicated what they buy together), whereas the clusters or segment we obtained with the Clustering model answered what kind of customers does the Retail Company have.

Finally, we applied deployment to the models, to give product recommendations to the customers in general, or to specific customers that belong to one cluster.

Section 8. Appendix

Troubleshooting

Back

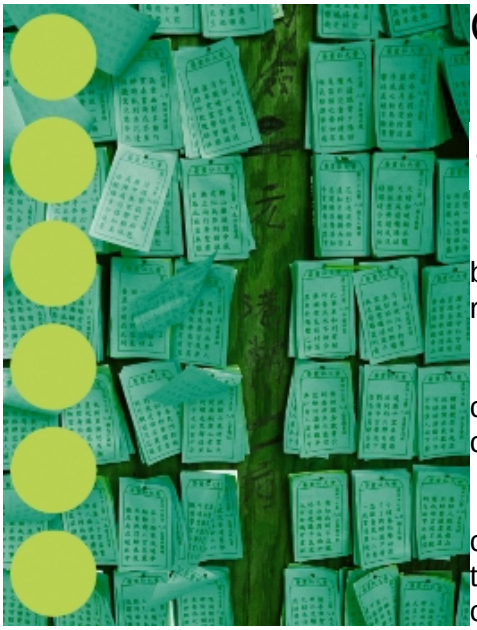
Starting the IBM DB2 Intelligent Miner Visualization:

Contact your DBA, to specify the correct URL and JDBC driver of your machine, if the one specified in this tutorial does not allow you to create a Database Profile.

Clustering Model:

In the data preparation, you cannot create nor update a table/view that is being showned with the IBM DB2 UDB. Close the window that shows the table/view first and then execute the SQL script.

If you cannot build the second model, close the IM Visualization window.



Glossary

Back

bodyIn a rule, the antecedent. For example, in the rule A and B ==> C, A and B are the body.

clusterA group of records with similar characteristics.

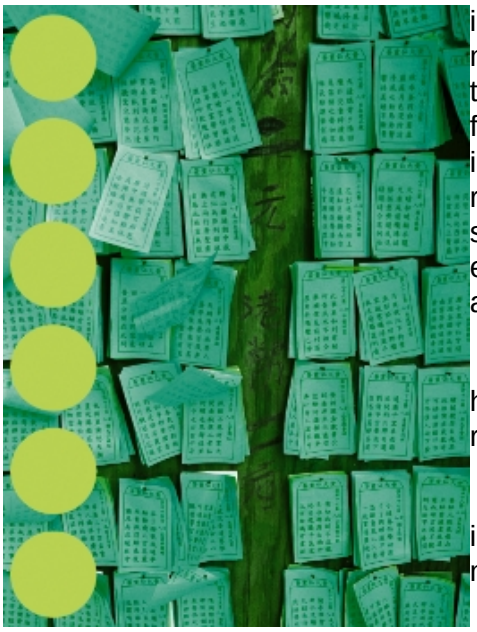
clustering mining function A data mining function that discovers sets of rows with common characteristics. These sets are known as clusters. Rows are possibly homogeneous inside a cluster, and possibly heterogeneous between two clusters.

confidence The percentage of transactions that contain all the items that make up a rule in the transactions that contain the antecedent of the rule. The confidence of the data mining association rule "A implies B" is 100 times the number of transactions that contain A and B divided by the number of transactions that contain A. In statistics, this could be written as:

Confidence = P(A and B) / P(A) = Support / P(A) = P(B/A)

cross-sellingThe strategy of pushing new products to current customers based on their past purchases. Cross-selling is designed to widen the customer's reliance on the company and decrease the likelihood of the customer switching to a competitor.

data miningThe process of discovering valuable, hidden facts, unknown information from a large amount of data. The data is analyzed without any preconceived expectation of the results. Data mining delivers knowledge that provides a deeper



insight into the data. Using a combination of machine learning, statistical analysis, modeling techniques and database technology, data mining finds patterns and subtle relationships in data and infers rules that allow the prediction of future results. Typical applications include market segmentation, customer profiling, fraud detection, evaluation of retail promotions, and credit risk analysis.

head In a rule, the consequent. For example, in the rule A and B ==> C, C is the head.

input data The tables or views you specify to be mined.

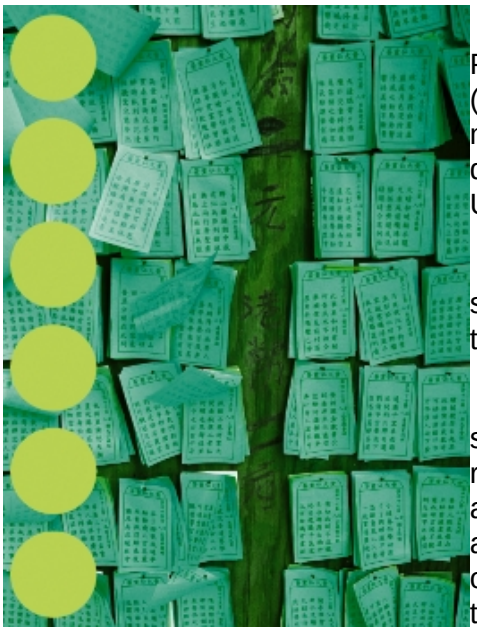
lift a measure of how the rule improves our ability to predict the head of the rule. For example, if the item A and B are purchased together (rule A ==> B), then if we have a Lift of 10 it means that the probability of finding B in those transactions that contain A as well, is 10 times higher than the probability of finding B in all the transactions (no matter if A was purchased or not). In statistics, this could be written as:

Lift = P(B/A) / P(B) = Confidence / P(B)

metadata In databases, data that describes data objects.

mining Synonym for analyzing or searching.

model An important function of data mining is the production of a model. A model can be descriptive or predictive. A descriptive model helps in understanding underlying processes or behaviour. For example, an association model describes consumer behaviour. A predictive model is an equation or set of rules that makes it possible to predict an unseen or unmeasured value from other, known values.



PMMLThe Predictive Model Markup Language (PMML) is a standard for exchanging data mining models in order to be able to deploy the models directly into relational databases, such as DB2 Universal Database.

scoreWhen applying the clustering mining function, the affinity of the customer record to his cluster

support The support of a data mining association rule is the percentage of transactions that contain all items of the rule. The support of the data mining association rule A ==> B is 100 times the number of transactions that contain A and B divided by the total number of transactions. In statistics, this could be written as:

Support = P(A and B)

GO BACK

Bibliography

This bibliography lists all related IBM publications that are relevant to the contents of this tutorial that might be useful for reference purposes.

Where appropriate, IBM publication numbers are given after the document title. This will assist you in finding the document online at the *IBM Publications Center*.

IBM Redbooks:

- *"Enhance Your Business Applications: Simple Integration of Advanced Data Mining Functions"* [ISBN 0738427799].
- *"Mining Your Own Business in Retail. Using DB2*

Intelligent Miner for Data" [ISBN 0738422940].

- *"Intelligent Miner for Data: Enhance your Business Intelligence."* [ISBN 0738413488].

Section 9. Feedback

Feedback

Please send us your feedback on this tutorial, specially on Troubleshooting. We look forward to hearing from you!

Colophon

This tutorial was written entirely in XML, using the developerWorks Toot-O-Matic tutorial generator. The open source Toot-O-Matic tool is an XSLT style sheet and several XSLT extension functions that convert an XML file into a number of HTML pages, a zip file, JPEG heading graphics, and two PDF files. Our ability to generate multiple text and binary formats from a single source file illustrates the power and flexibility of XML. (It also saves our production team a great deal of time and effort.)

You can get the source code for the Toot-O-Matic at www6.software.ibm.com/dl/devworks/dw-tootomatic-p. The tutorial Building tutorials with the Toot-O-Matic demonstrates how to use the Toot-O-Matic to create your own tutorials. developerWorks also hosts a forum devoted to the Toot-O-Matic; it's available at www-105.ibm.com/developerworks/xml_df.nsf/AllViewTemplate?OpenForm&RestrictToCategory=11. We'd love to know what you think about the tool.